Structural Identifiability of large Systems Biology models

Dominique Joubert

STRUCTURAL IDENTIFIABILITY OF LARGE SYSTEMS BIOLOGY MODELS

Dominique Joubert

Thesis committee

Promotor
Prof. Dr J. Molenaar
Professor of Applied Mathematics,
Wageningen University & Research

Co-promotor
Dr J. D. Stigter
Assistant Professor, Mathematical and Statistical Methods Group
Wageningen University & Research

Other members
Prof. Dr K. J. Keesman, Wageningen University & Research
Dr C. Fleck, ETH Zürich, Switzerland
Prof. Dr N. A. W. van Riel, University of Amsterdam / Eindhoven University of Technology
Prof. Dr P. M. J. van den Hof, Eindhoven University of Technology

This research was conducted under the auspices of the Graduate School of Production Ecology and Resource Conservation (PE&RC)

STRUCTURAL IDENTIFIABILITY OF LARGE SYSTEMS BIOLOGY MODELS

Dominique Joubert

Thesis

submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. Dr A. P. J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Monday 27 January 2020
at 4 p.m. in the Aula.

Dominique Joubert
Structural Identifiability of large Systems Biology models,
173 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2020)
With references, with summary in English

ISBN: 978-94-6395-226-2
DOI https://doi.org/10.18174/507603

ABSTRACT

A fundamental principle of systems biology is its perpetual need for new technologies that can solve challenging biological questions. This precept will continue to drive the development of novel analytical tools. The virtuous cycle of biological progress can therefore only exist when experts from different disciplines including biology, chemistry, computer science, engineering, mathematics, and medicine collaborate. General opinion is however that one of the challenges facing the systems biology community is the lag in the development of such technologies.

The topic of structural identifiability in particular has been of interest to the systems biology community. This is because researchers in this field often face experimental limitations. These limitations, combined with the fact that systems biology models can contain vast numbers of unknown parameters, necessitate an identifiablity analysis. In reality, analysing the structural identifiability of systems biology models, even when they contain only a few states and system parameters, may be challenging. As these models increase in size and complexity, this difficulty is exasperated, and one becomes limited to only a few methods capable of analysing large ordinary differential equation models. In this thesis I study the use of a computationally efficient algorithm, well suited to the analysis of large models, in the model development process.

The three related objectives of this thesis are: 1) develop an accurate method to asses the structural identifiability of large possibly nonlinear ordinary differential models, 2) implement this method in the preliminary design of experiments, and 3) use the method to address the topic of structural unidentifiability.

To improve the method's accuracy, I systematically study the role of individual factors, such as the number of experimentally measured sensors, on the sharpness of results. Based on the findings, I propose measures that can improve numerical accuracy.

To address the second objective, I introduce an iterative identifiability algorithm that can determine minimal sets of outputs that need to be measured to ensure a model's local structural identifiability. I also illustrate how one could potentially reduce the computational demand of the algorithm, enabling a user to detect minimal output sets of large ordinary differential equation models within minutes.

For the last objective, I investigate the role of initial conditions in a model's structural unidentifiability. I show that the method can detect problematic values for large ordinary differential equation models. I illustrate its role in reinstating the local structural identifiability of a model by identifying problematic initial conditions.

I also show that the method can provide theoretical suggestions for the reparameterisation of structurally unidentifiable models. The novelty of this work is that the algorithm allows for unknown initial conditions to be parameterised and accordingly, repameterisations requiring the transformation of states, associated with unidentifiable initial conditions, can easily be obtained. The computational efficiency of the method allows for the reparameterisation of large ordinary differential equation models in particular.

To conclude, in this thesis I introduce an method that can be used during the model development process in an array of useful applications. These include: 1) determining minimal output sets, 2) reparameterising structurally unidentifiable models and 3) detecting problematic initial conditions. Each of these application can be implemented *before any experiments are conducted* and can play a potential role in the optimisation of the modelling process.

Keywords: Structural identifiability, minimal output sets, experimental design, reparameterization, ordinary differential equations, nonlinear systems, systems biology, singular value decomposition.

CONTENTS

1

GENERAL INTRODUCTION

Dominique JOUBERT

The lack of real contact between mathematics and biology is either a tragedy, a scandal or a challenge, it is hard to decide which.

(Gian-Carlo Rota, 1932-1999)

1

OUR aspirations for in depth knowledge continue to drive us toward the interface between different scientific disciplines. This is evident in fields such as finance, which requires both financial and mathematical expertise, and also in systems biology, where biological knowledge meets mathematical topics such as optimisation and integration. These "knowledge interfaces" allow for rapid scientific progress and in the future, will be the norm rather than the exception.

The work covered in this thesis is located at such an interface, with mathematics and engineering complementing biology. The aim is to provide model developers with practical tools that can be used during both experimental design and parameter estimation. More concisely, we study a fast algorithm that can be used to analyse a model's local structural identifiability and due to its computational efficiency, we extend its use to the design of experiments.

In this general introduction, the reader is guided through a series of topics that culminates in the discussion on the main subject of this thesis, structural identifiability. These include a general discussion on the subject of systems biology, the importance of dynamic modelling and a glimpse into the iterative process involved in creating a reliable dynamic model. The topic of identifiability is formally introduced after which selected challenges facing the structural identifiability (SI) community are highlighted. These lead to the formal definition of the objectives of this thesis.

1.1. SYSTEMS BIOLOGY

Although the study of life dates back millennia, our understanding of the different mechanisms operating within living organisms was limited until the 1950s. During this period, the field of molecular biology began to offer more detailed descriptions of networks between interacting molecules. These descriptions were made possible due to the tiny spacial scales at which molecular processes could be observed during experiments, leading to a reductionist modelling approach, where scientists aimed to describe individual molecules [1].

Improvements in experimental techniques at the beginning of the 21^{st} century, set the stage for a shift in this modelling perspective. The emergence of high-throughput approaches allowed experimental researchers to observe the behaviour of large groups of distinct molecular species simultaneously. A catalyst for these developments was the sequencing of the human genome, of which the first draft appeared in 2000. Consequently, modern research is driven by experiments that reveal the behaviour of entire molecular systems, describing the interrelations and interactions within the contexts of time, space, and physiology [2]. This is known as systems biology [1].

Today, systems biology is a well established and interdisciplinary field that draws from various scientific disciplines including mathematics, bioinformatics, computer science, and engineering [3–6]. A general opinion is that one of the main challenges in systems biology research, is the development of technologies required to analyse models. Examples of such technologies are mathematical theories and algorithms that allow for the efficient analysis of nonlinear models, and sophisticated software that facilitates the examination and integration of large amounts of data [2].

1.2. DYNAMIC MATHEMATICAL MODELLING

Modelling and simulation enable us to integrate and summarise information, perform *in silico* experiments, and generate predictions and hypotheses that can increase our understanding of complex systems [7]. Despite the fact that these simulations can never replace laboratory experiments, they are useful. For example, they can be used to examine a system's behaviour in ways that would not be attainable in a lab. In addition, they can be carried out quickly and in contrast with actual experiments, incur no significant costs [1]. Ordinary differential equation (ODE) models enable us to describe most dynamic systems. Due to advances in experimental procedures, high quality *time-series* data can be collected from which these models can be calibrated. Consequently, ODE models have become the models of choice in various biomedical research fields [8] and as such, these models will be analysed in this thesis.

1.2.1. ORDINARY DIFFERENTIAL EQUATIONS

ODE models are used in disciplines ranging from electrical and chemical engineering, to biology and medicine. The majority of the models analysed in this thesis comprise the following components: state variables and their initial conditions, system parameters, and outputs. Selected models in chapter 5 also contain inputs. *State variables* describe the states of individual components within a model. The collection of states describes the condition of the system as a whole at any given time. *Parameters* characterise interactions among the different states and their values can either be known or unknown. Given that parameters can usually not be measured directly, values for unknown parameters need to be inferred from observed data and this is known as parameter estimation or calibration. Experimentally measurable states/sensors are defined as *outputs*. An *input* is the controlled part of a system that helps it achieve a specific output.

Once values for the unknown system parameters have been calculated, questions regarding their accuracy and whether their values are unique may arise. The issue of parameter uniqueness is referred to as *identifiability*. Because the quantitative descriptions of molecular interactions typically invoke the laws of physics and chemistry and many of these relationships are nonlinear, nearly all systems are nonlinear in nature [9]. This nonlinearity, in conjunction with the increasing size of modern models, significantly complicates the calibration of ODE models.

1.2.2. MODELLING PROCESS

The modelling process necessitates a critical evaluation of the underlying mechanisms of a system. A model is the construct of our current understanding of a system and its results can be useful in numerous ways. For example, they can help in the design of experiments by indicating promising avenues for investigation. They may also reveal inconsistencies between our understanding of a system (encapsulated in the model) and experimental observations. The presence of these inconsistencies is one of the main advantageous of modelling in the sense that outcomes that do not reflect experimental observations can be regarded as a falsification of the original hypotheses and so, our understanding of certain mechanisms may be updated. This leads to the refinement of hypotheses and an updated model structure, which can in turn be tested against additional experiments. This iterative process, the so-called "modelling" or virtuous cycle,

results in the continuous improvement of a model [1].

Figure 1.1 is a representation of such a cycle. The topics covered in this thesis, and where they fit into this cycle are also indicated. In the coming chapters, the importance of a preliminary evaluation of one's model is emphasised and this leads us to a formal discussion of *a priori* or structural identifiability analysis.

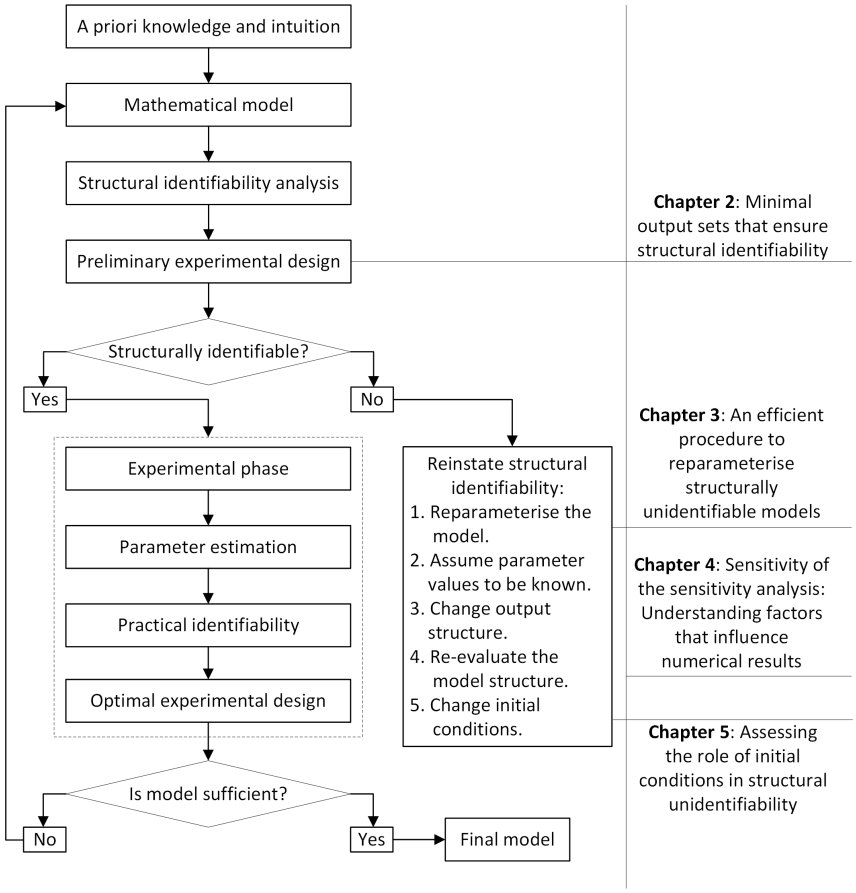


Figure 1.1: **Model development process.** The so-called "modelling" or virtuous cycle results in the continuous improvement of a model. The stages of the cycle are shown on the left. The different topics covered in this thesis, and where they fit into this cycle are indicated on the right.

1.3. STRUCTURAL IDENTIFIABILITY

The question regarding the uniqueness of inferred parameter estimates can be answered by performing an identifiability analysis. A model's identifiability is determined by either its structure, in which case one refers to *a priori* or structural identifiability, or the richness of experimental data, referred to as practical identifiability.

The topic of structural identifiability has been of particular interest to the systems biology community. This is because researchers in this field often face experimental limitations [10]. These limitations, combined with the fact that systems biology models can contain vast numbers of unknown parameters, necessitate an identifiablity analysis at an early stage. From figure 1.1 one sees that an *a priori* or structural analysis of a model can be done before entering the experimental phase. This step is vital in ensuring that accurate parameter estimates can be calculated.

In reality, analysing the structural identifiability of systems biology models, even when they contain only a few states and system parameters, may be challenging. As these models increase in size and complexity, this difficulty is exasperated and one becomes limited to only a few methods capable of analysing large ODE models.

1.3.1. A SIMPLE BIOLOGICAL PROBLEM

Consider an experimental setup involving a batch reactor filled with a homogeneous mixture of Lactobacillus cells, of concentration l, and nutrients, of concentration n. The aim is to develop a model that describes the growth of these bacterial cells over the course of time. Accordingly, we use an ODE system.

Both the nutrient and Lactobacillus cell concentrations are modelled as states with initial values, $n(0)$ and $l(0)$, respectively. We assume that these initial concentrations are unknown and as such, they can be regarded as additional parameters that need to be inferred from experimental data. Furthermore, we use the well know Monod equation to describe the concentrations of both the cells and the nutrients. The model therefore has 5 unknown parameters in total. Three system parameters, μ, ϕ, and γ, related to the growth dynamics, and the 2 unknown initial conditions, $l(0)$ and $n(0)$. The measured output, denoted as $y = n$, indicates that we can only measure the nutrient concentration experimentally. This model can be written in the standard state-space form:

$$\dot{l}(t) = l(t)\frac{\mu n(t)}{\phi + n(t)}, \tag{1.1}$$

$$\dot{n}(t) = -\frac{1}{\gamma} l(t)\frac{\mu n(t)}{\phi + n(t)}, \tag{1.2}$$

$$l(0) = l_0, n(0) = n_0, \tag{1.3}$$

$$y(t) = n(t). \tag{1.4}$$

The evaluation of whether or not it is possible to find unique estimates for the 2 initial conditions, $l(0)$ and $n(0)$, given the properties of the output structure, is known as an *observability analysis*. This concept can be extended to allow for the analysis of whether unique estimates for all the unknown parameters, $\boldsymbol{\theta} = [\mu, \phi, \gamma, l(0), n(0)]$, can be found from the measured output, y. This is known as a *structural identifiability* (SI) analysis.

Due to its structural properties, SI can be analysed without the availability of any experimental data. Knowing *a priori* whether there is any chance of uniquely estimating all unknown parameters may potentially save on both experimental time and expenses. Therefore, the structural identifiability of a model should ideally be checked before collecting data [11].

1

1.3.2. DEFINITIONS

In this section, formal definitions for some of the terms in this thesis are given.

Definition 1: **Indistinguishability**. Let M be a model with state x and measurable output y. Let $y_{x0}(t)$ denote the time evolution of the model output when started from an initial state x_0 at t_0. Two states x_1 and x_2 are indistinguishable if $y_{x1}(t) = y_{x2}(t)$ for all $t \geq t_0$. The set of states that are indistinguishable from x_1 is denoted by $I(x_1)$ [10].

Definition 2: **Observability**. A model M is observable at x_0 if $I(x_0) = x_0$ [12]. This property implies that the initial state of the system can be deduced from observing the output y.

To understand how one assesses a system's observability, first consider the following linear time invariant (LTI) model, M_{LTI}:

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{x}(t), \tag{1.5}$$

$$\boldsymbol{y}(t) = \boldsymbol{C}(\boldsymbol{\theta})x(t), \tag{1.6}$$

$$\boldsymbol{x_0} = \boldsymbol{x}(t_0, \boldsymbol{\theta}), \tag{1.7}$$

where $\boldsymbol{\theta} \in R^p$ is the parameter vector, $\boldsymbol{x}(t) \in R^n$ the state vector, and $\boldsymbol{y}(t) \in R^m$ the output vector. $\boldsymbol{A}(\boldsymbol{\theta})$ and $\boldsymbol{C}(\boldsymbol{\theta})$ are constant matrices of dimensions $n \times n$ and $m \times n$, respectively.

One way of assessing the observability of M_{LTI} is to check for the so-called observability rank condition (ORC). The rank of the observability matrix, defined as $\mathcal{O} = (\boldsymbol{C}|\boldsymbol{C.A}|\boldsymbol{C.A}^2|\ \boldsymbol{C.A}^3|\ldots|\boldsymbol{C.A}^{(n-1)}|)^T$, is calculated and if the rank$(\mathcal{O}) = n$, model M_{LTI} is observable. This is summarised in the following theorem:

Theorem 1: **Linear Observability Rank Condition**. For the LTI model M_{LTI}, defined in 1.5-1.7, a necessary and sufficient condition for M_{LTI} to be classified as observable is that rank$(\mathcal{O}) = n$ [13], where:

$$\mathcal{O} = \begin{pmatrix} \boldsymbol{C} \\ \boldsymbol{C.A} \\ \boldsymbol{C.A}^2 \\ \vdots \\ \boldsymbol{C.A}^{n-1} \end{pmatrix} = \frac{\partial}{\partial \boldsymbol{x}} \begin{pmatrix} \boldsymbol{y}(t) \\ \dot{\boldsymbol{y}}(t) \\ \ddot{\boldsymbol{y}}(t) \\ \vdots \\ \boldsymbol{y}^{n-1}(t) \end{pmatrix}. \tag{1.8}$$

Now consider a nonlinear system, M_{NL}:

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{f}(t, \boldsymbol{x}(t), \boldsymbol{\theta}), \tag{1.9}$$

$$\boldsymbol{y}(t) = \boldsymbol{h}(\boldsymbol{x}(t), \boldsymbol{\theta}), \tag{1.10}$$

$$\boldsymbol{x_0} = \boldsymbol{x}(t_0, \boldsymbol{\theta}). \tag{1.11}$$

State variables are contained in the vector $\boldsymbol{x}(t)$ (dim $\boldsymbol{x} = n$), and model parameters in vector $\boldsymbol{\theta}$ (dim $\boldsymbol{\theta} = p$). The output or measured states/sensors are contained in vector $\boldsymbol{y}(t)$ (dim $\boldsymbol{y} = m$). Function $\boldsymbol{f}$ denotes a dynamic model structure and $\boldsymbol{h}$ is the output or observation function and both these are assumed to be analytic functions and can either be rational or irrational.

Obtaining the observability matrix for nonlinear models requires the computation of Lie derivatives. A Lie derivative, denoted as $\mathscr{L}_{\boldsymbol{f}}\boldsymbol{h}(\boldsymbol{x})$, is the directional derivative of the smooth function, $\boldsymbol{h}(\boldsymbol{x})$, with respect to the vector field, $\boldsymbol{f}(\boldsymbol{x})$, which describes the model dynamics. It is defined as [10]:

$$\mathscr{L}_{\boldsymbol{f}}\boldsymbol{h}(\boldsymbol{x}) = \frac{\partial \boldsymbol{h}(\boldsymbol{x})}{\partial \boldsymbol{x}}\boldsymbol{f}(\boldsymbol{x}), \tag{1.12}$$

with higher order derivatives computed consecutively as:

$$\mathscr{L}^{i}_{\boldsymbol{f}}\boldsymbol{h}(\boldsymbol{x}) = \frac{\partial \mathscr{L}^{i-1}_{\boldsymbol{f}}\boldsymbol{h}(\boldsymbol{x})}{\partial \boldsymbol{x}}\boldsymbol{f}(\boldsymbol{x}). \tag{1.13}$$

Accordingly, the nonlinear observability matrix can be computed by calculating successive Lie derivatives:

$$\mathscr{O}(\boldsymbol{x}(0)) = \begin{pmatrix} \frac{\partial}{\partial \boldsymbol{x}}\boldsymbol{y}(t) \\ \frac{\partial}{\partial \boldsymbol{x}}\dot{\boldsymbol{y}}(t) \\ \frac{\partial}{\partial \boldsymbol{x}}\ddot{\boldsymbol{y}}(t) \\ \vdots \\ \frac{\partial}{\partial \boldsymbol{x}}\boldsymbol{y}^{n-1}(t) \end{pmatrix} = \begin{pmatrix} \frac{\partial \boldsymbol{h}(\boldsymbol{x})}{\partial \boldsymbol{x}} \\ \frac{\partial}{\partial \boldsymbol{x}}(\mathscr{L}_{\boldsymbol{f}}\boldsymbol{h}(\boldsymbol{x})) \\ \frac{\partial}{\partial \boldsymbol{x}}(\mathscr{L}^{2}_{\boldsymbol{f}}\boldsymbol{h}(\boldsymbol{x})) \\ \vdots \\ \frac{\partial}{\partial \boldsymbol{x}}(\mathscr{L}^{n-1}_{\boldsymbol{f}}\boldsymbol{h}(\boldsymbol{x})) \end{pmatrix}. \tag{1.14}$$

Theorem 2: **Nonlinear Observability Rank Condition**. For a nonlinear model M_{NL}, as defined in 1.9-1.11, if the rank of the observability matrix, as defined in 1.14 is n, then M_{NL} is locally observable around $\boldsymbol{x}(0)$ [14, 15].

Let us expand on the concept of observability by including system parameters to our analyses. We start with a formal definition of identifiability:

Definition 3: **Identifiability**. The dynamic system defined in 1.9-1.11 is identifiable if all the parameters defined in $\boldsymbol{\theta}$ can uniquely be determined from the measurable output $\boldsymbol{y}(t)$, assumed to be noise-free and continuous in time [16]. Otherwise, it is classified as unidentifiable [8].

Ljung and Glad distinguish between locally identifiable and globally identifiable systems [17].

Definition 4: **Local structural identifiability**. A model M_{NL}, given in 1.9-1.11, is structurally locally identifiable (s.l.i.) if for almost any parameter vector $\boldsymbol{\theta}^* \in \mathbb{R}^p$, there is a neighbourhood $\mathscr{N}(\boldsymbol{\theta}^*)$ such that the following property holds [17]:

$$\boldsymbol{\theta} \in \mathscr{N}(\boldsymbol{\theta}^*) \text{ and } \boldsymbol{h}(\boldsymbol{x}(t),\boldsymbol{\theta}) = \boldsymbol{h}(\boldsymbol{x}(t),\boldsymbol{\theta}^*) \Rightarrow \boldsymbol{\theta} = \boldsymbol{\theta}^*, \tag{1.15}$$

for $t \geq t_0$.

Definition 5: **Globally structural identifiability**. A model M_{NL}, given in 1.9-1.11, is structurally globally identifiable (s.g.i.) if all its parameters can be uniquely determined

from the system output, that is for almost any parameter vector $\boldsymbol{\theta}^* \in \mathbb{R}^p$, the following property holds [17]:

$$h(\boldsymbol{x}(t),\boldsymbol{\theta}) = h(\boldsymbol{x}(t),\boldsymbol{\theta}^*) \Rightarrow \boldsymbol{\theta} = \boldsymbol{\theta}^*, \tag{1.16}$$

for $t \geq t_0$. At this point, one arrives at a junction in terms of how to go about combining the concepts of structural identifiability and observability. On the one hand, parameters can be regarded as special model states with zero dynamics, in which case the nonlinear observability rank condition would reveal the observability of both initial conditions of states and system parameters [10, 12, 18–20]. Alternatively, one can parameterise model states and regard them as additional system parameters, in which case a structural identifiability analysis would reveal the identifiability of both system parameters and the initial conditions of model states [21]. In this thesis, we adopt the latter approach and describe this process in great detail in the coming chapters. For completion, we discuss the concept of augmenting the state vector in this general introduction.

By regarding parameters as special model states with zero dynamics, the state vector can be augmented as follows: $\tilde{\boldsymbol{x}} = [\boldsymbol{x},\boldsymbol{\theta}]$ $(dim(\tilde{\boldsymbol{x}}) = n+p)$. This allows us to define structural identifiability as a special case of observability [10]. Because the nonlinear observability theorem holds for local conditions only, one can only analyse local structural identifiablity using this approach. The augmented observability matrix is defined as [19]:

$$\mathscr{O}(\tilde{\boldsymbol{x}}(0)) = \mathscr{O}(\boldsymbol{x}_0,\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial}{\partial \tilde{\boldsymbol{x}}}\boldsymbol{y}(t) \\ \frac{\partial}{\partial \tilde{\boldsymbol{x}}}\dot{\boldsymbol{y}}(t) \\ \frac{\partial}{\partial \tilde{\boldsymbol{x}}}\ddot{\boldsymbol{y}}(t) \\ \vdots \\ \frac{\partial}{\partial \tilde{\boldsymbol{x}}}\boldsymbol{y}^{n+p-1}(t) \end{pmatrix} = \begin{pmatrix} \frac{\partial \boldsymbol{h}(\tilde{\boldsymbol{x}})}{\partial \tilde{\boldsymbol{x}}} \\ \frac{\partial}{\partial \tilde{\boldsymbol{x}}}(\mathscr{L}_f \boldsymbol{h}(\tilde{\boldsymbol{x}})) \\ \frac{\partial}{\partial \tilde{\boldsymbol{x}}}(\mathscr{L}_f^2 \boldsymbol{h}(\tilde{\boldsymbol{x}})) \\ \vdots \\ \frac{\partial}{\partial \tilde{\boldsymbol{x}}}(\mathscr{L}_f^{n+p-1} \boldsymbol{h}(\tilde{\boldsymbol{x}})) \end{pmatrix}. \tag{1.17}$$

This matrix can be calculated symbolically using software implemented in e.g. Mathematica.

Theorem 3: **Nonlinear Identifiability-Observability Rank Condition**. For a nonlinear model M_{NL} as defined in 1.9-1.11, if the augmented observability matrix, as defined in 1.17 has rank $n+p$, then M_{NL} is locally observable and structurally identifiable in a neighbourhood $\mathscr{N}(\tilde{\boldsymbol{x}}(0))$ of $\tilde{\boldsymbol{x}}(0)$ [10].

If 1.17 is of full rank, all parameters and initial conditions are structurally locally identifiable and if this matrix is rank deficient, the parameter space is divided between structurally identifiable and structurally unidentifiable parameters.

1.3.3. HISTORY AND CURRENT TRENDS

The concept of observability for linear time-invariant systems was introduced by Kalman in 1960 [13, 22]. The demand for analysing nonlinear models soon led to the develop-

ment of several methods [14, 23–26]. Addressing the interest in parameter identifiability, Bellman and Åström introduced the concept of structural identifiability for linear models in 1970 [27]. They suggested using a *Laplace transform* to analyse models. In 1982 Tunali and Tarn extended this concept for nonlinear models [20]. Even though there are a number of methods available for the analysis of nonlinear models, no single method is amenable to all models. Refer to [7, 8] for a review of the methods available. Structural identifiablity methods can in general be divided into the following classes:

- *Series expansion*: This includes approaches using Taylor series expansion and the power series method for linear models [28]. Taylor series methods are described in papers by Margaria *et. al.* [29] and Vajda [30]. Authors using the power series approach include: Pohjanpalo [28], Walter and Pronzato [16] and Chis *et. al.* with the package GENSII [31, 32]. August and Papachristodoulou use symbolic calculations to calculate an augmented observability matrix for the analysis of local identifiability for smaller systems [19]. The package STRIKE-GOLDD is in principle suited for large models [33].

- *Differential algebra (Global structural identifiability)*: These approaches can be divided in two groups. One group treats parameters as functions with zero derivatives and uses differential elimination [34]. Ljung and Glad apply this to small systems [17]. The second group treats parameters as elements of the field of coefficients and produce so-called input-output equations. This approach is followed in the packages COMBOS [35] and DAISY [36, 37]. Other contributions include those from Diop and Fliess [38], Carra [39], and Ollivier [40].

- *Differential geometry (Local structural identifiability)*: This approach is implemented by, amongst others, Isidori [41], Sontag [42] and Vidyasagar [15]. It is based on the Inverse Function Theorem from calculus.

- *Semi-numerical*: Karlsson *et. al.* [43] and Sedoglavic [18] implement semi-numerical methods in their papers.

- *Numerical (Local structural identifiability)*: Numerical methods are used by amongst others Raue *et. al.* [44]. For rational systems, the Exact Arithmetic Rank (EAR) method is a numerical rank calculating method [43], based on the algorithm introduced by Sedoglavic in 2002 [18]. Other authors also use the correlation method [45], principle component analysis (PCA) [46], the orthogonal method [47], and the eigenvalue method [48] to analyse the calculated sensitivity matrix.

- *Other*: A direct test, both analytical [49] and numerical [50]. Walter and Lecourtier introduced a similarity transform method for linear models [51]. Vadja and Rabitz [52], Vajda *et. al.* [53] and Chappel and Godfrey [54] extended this similarity transform approach to nonlinear models. This method requires the system to be both observable and controllable. Xia and Moog proposed another method based on the Implicit Function Theorem [55]. Wu *et al.* [56] further extended this method. Stigter and Molenaar introduced a hybrid method incorporating a sensitivity based method [57] that uses singular value decomposition (SVD) to calculate

the rank of a sensitivity matrix, and a subsequent series expansion method based on Pohjanpalo's power series expansion.

The hybrid method of Stigter and Molenaar [57] is used in this thesis. It has been chosen due to its computational efficiency, allowing for it to be used in applications complementary to straight forward identifiability analyses. In addition, its numerical results are easy to interpret and decrease the computational demand of the subsequent symbolic calculations. This attribute makes the method well-suited for the analysis of large systems biology type of models.

1.4. PROBLEM STATEMENT

Our problem statement is defined by looking toward the future. Given the growing interest in the field of systems biology, increasing amounts of generated data, and the never-ending consolidation of our knowledge, it is evident that in the future, models will continue to increase in size and complexity. As Kolczyk recently stated, "Answering new and even more complex biomedical questions require models of complete cells, organs or even organisms" [58]. This trend necessitates the development of software that is capable of handling such systems and this remains a challenging open problem. As Ghosh *et. al.* summarises it: "Understanding complex biological systems requires extensive support from software tools. Such tools are needed at each step of a systems biology computational workflow, which typically consists of data handling, network inference, deep curation, dynamical simulation and model analysis. In addition, there are now efforts to develop integrated software platforms, so that tools that are used at different stages of the workflow and by different researchers can easily be used together" [59]. In the identifiability context, the aim is to facilitate the accurate analysis of large systems. This leads to the definition of the general aim of this thesis: **Provide an easy to use structural identifiability method that can be applied to a wide range of systems**. It should ideally be capable of analysing both rational and irrational model functions.

In [10] it is stated: "If a set of parameters is found to be structurally unidentifiable, a question naturally arises: is it possible to reformulate the model by combining such parameters in an identifiable quantity? The answer to this question entails characterising the way the structurally unidentifiable parameters are correlated. Many methods for structural identifiability analysis are capable of addressing this problem to a certain extent; however, no generally applicable and automatic procedure exists". This leads to the definition of our second goal: in the event of structural unidentifiability, model developers should be well-informed as to the different avenues available to reinstate a particular model's identifiability. **An easy to implement method should assist model developers with the reparameterisation of their structurally unidentifiable models**. Ideally, it should be capable of identifying and eliminating the redundant parameters of large ODE models.

The third goal is closely related to the notion of giving model developers different options when addressing structural unidentifiability. "It happens frequently in the global identifiability applications that the property holds only generically, i.e. except for a "thin" set of initial conditions. In these situations the system is (incorrectly but forgivably) nevertheless declared to be (global) identifiable, excluding certain subsets of initial states"

[60]. **Our method should be able to detect certain problematic initial conditions that, if changed, would reinstate a model's structural identifiability**.

Given that a model can be analysed prior to the experimental phase of a model's development, structural identifiability algorithms can in principle be implemented in a early experimental design capacity. This may entail detecting which outputs should be measured to ensure a model's structural identifiability. Anguelova has the following to say regarding the minimal output set problem: "This problem has not received much attention previously, possibly due to the difficulty of testing structural identifiability in the first place" [61]. Given the increase in the number of experiments conducted and the costs incurred during such experiments, there is a need for optimally-designed experiments. This includes factors such as how many experimental measurements should be taken, at which time points should these be taken and **which sensors should be measured to ensure a model's structural identifiability?**

Finally, light should be shed on the numerical methods used for local structural identifiability analyses. Evans states that "Numerical analysis is heavily dependent on notional values for the parameters (that are to be estimated), and involves applying a sampling rate to the output. These results are therefore affected by a number of factors that one would wish to understand the individual effect of - for example, is a model overparameterised regardless of the number and timing of samples taken" [62]. To summarise, **a better understanding of the factors that influence numerical identifiability results is required**.

1.5. OBJECTIVES AND THESIS OUTLINE

Given: 1) the open research questions that exist within the identifiability community, 2) the lack of interaction between different scientific disciplines, and 3) the limited amount of software tools capable of addressing the problems that arise during model development, the following objectives have been earmarked for this thesis. Figure 1.1 shows where they fit into the model development process.

1. Identifiability software is implemented in the preliminary design of experiments. Here, the *minimal sets of outputs* that need to be measured to ensure a model's structural identifiability are determined. This is covered in **Chapter 2**: *Determining minimal output sets that ensure structural identifiability.*

2. Address the topic of *structural unidentifiability.* The different options available to reinstate structural identifiability are discussed and the process of model reparameterisation is shown in detail. In addition, model reparameterisation involving state transformations of large ODE models is also shown. **Chapter 3**: *An efficient procedure to reparameterise structurally unidentifiable models.*

3. Identify the key *factors that influence the numerical structural identifiablity results.* The motivation for this objective is that a good understanding of these factors will enable the us to apply our method to a wide range of models. **Chapter 4**: *Sensitivity of the sensitivity analysis: Understanding factors that influence numerical results.*

4. Continuing on the theme of reinstating structural identifiability, *we introduce a fast method that can accurately detect sets of problematic initial conditions of large ODE models.* Here, the effect of initial conditions on a model's structural identifiability is shown. **Chapter 5**: *Assessing the role of initial conditions in the local structural identifiability of large nonlinear dynamical models.*

5. A general discussion of the achieved results is given in this chapter. Conclusions are drawn and finally, possible future applications of the method are discussed. **Chapter 6**: *General discussion.*

REFERENCES

[1] B. Ingalls, *Mathematical Modeling in Systems Biology: An Introduction* (The MIT Press, Cambridge, MA., 2013).

[2] C. Stadtländer, *Systems biology: mathematical modeling and model analysis,* Journal of Biological Dynamics **12**, 11 (2018).

[3] A. Aderem, *Systems biology: Its practice and challenges,* Cell **121**, 511 (2005).

[4] J. Green, A. Hastings, P. Arzberger, F. Ayala, K. Cottingham, K. Cuddington, F. Davis, J. Dunne, M. Fortin, L. Gerber, and M. Neubert, *Complexity in ecology and conversation: Mathematical, statistical, and computational challenges,* BioScience **55**, 501 (2005).

[5] B. Karahalil, *Overview of systems biology and omics technologies,* Curr. Med. Chem. **23**, 4221 (2016).

[6] J. Vera, X. Lai, U. Schmitz, and O. Wolkenhauer, *Microrna-regulated networks: The perfect storm for classical molecular biology, the ideal scenario for systems biology,* in *MicroRNA Cancer Regulation. Advances in Experimental Medicine and Biology, vol 774*, edited by V. J. Schmitz U., Wolkenhauer O. (Springer, Dordrecht, 2013).

[7] B. J. Chis, O.T. and E. Balsa-Canto, *Structural identifiability of systems biology models: A critical comparison of methods,* PLoS ONE **6**, e27755 (2011).

[8] H. Miao, X. Xia, A. Perelson, and H. Wu, *On identifiability of nonlinear ode models and applications in viral dynamics,* SIAM Rev Soc Ind Appl Math **53**, 3 (2011).

[9] J. de Canete, C. Galindo, and I. I. Garcia-Moral, *System Engineering and Automation: An Interactive Educational Approach* (Berlin: Springer-Verlag, 2011).

[10] A. Villaverde, *Observability and structural identifiability of nonlinear biological systems,* Complexity **8497093** (2019), 10.1155/2019/8497093.

[11] M. Saccomani and K. Thomaseth, *The union between structural and practical identifiability makes strength in reducing oncological model complexity: A case study,* Complexity **2380650** (2018), 10.1155/2018/2380650.

[12] M. Anguelova, *Observability and identifiability of nonlinear systems with applications in biology*, Ph.D. thesis, Göteborg University (2007).

[13] R. Kalman, ed., *On the general theory of control systems* (1960) in Proc. 1st IFAC World Congress.

[14] R. Hermann and A. Krener, *Nonlinear controllability and observability,* IEEE Transactions on Automatic Control **22**, 728 (1977).

[15] M. Vidyasagar, *Nonlinear systems analysis* (Prentice Hall, Englewood Cliffs, NJ, 1993).

[16] E. Walter and L. Pronzato, *Identification of parametric models from experimental data,* in *Communications and Control Engineering Series* (Springer, London, UK, 1997).

[17] L. Ljung and T. Glad, *On global identifiability for arbitrary model parametrizations,* Automatica **30**, 265 (1994).

[18] A. Sedoglavic, *A probabilistic algorithm to test local algebraic observability in polynomial time,* Journal of Symbolic Computation **33**, 735 (2002).

[19] E. August and A. Papachristodoulou, *A new computational tool for establishing model parameter identifiability,* Journal of Computational Biology **6**, 875 (2009).

[20] E. T. Tunali and T. J. Tarn, *New results for identifiability of nonlinear systems,* IEEE Transactions on Automatic Control **32**, 146 (1987).

[21] J. D. Stigter and R. L. M. Peeters, *On a geometric approach to the structural identifiability problem and its application in a water quality case study,* in *2007 European Control Conference (ECC)* (2007) pp. 3450–3456.

[22] R. Kalman, *Contributions to the theory of optimal control,* Boletin de la Sociedad Matematica Mexicana **5**, 102 (1960).

[23] E. Griffith and K. Kumar, *On the observability of nonlinear systems,* Journal of Mathematical Analysis and Applications **35**, 135–147 (1971).

[24] Y. Kostyukovskii, *Simple conditions of observability of nonlinear controlled systems,* Autom. Remote Control **29**, 1575–1584 (1968).

[25] S. Kou, D. Elliott, and T. Tarn, *Observability of nonlinear systems,* Information and Control **22**, 89 (1973).

[26] H. Sussmann and V. Jurdjevic, *Controllability of nonlinear systems,* J. Differ. Equations. **12**, 95 (1972).

[27] R. Bellman and K. J. Aström, *On structural identifiability.* Mathematical Biosciences **7**, 329 (1970).

[28] H. Pohjanpalo, *Systems identifiability based on the power series expansion of the solution,* Mathematical Biosciences **41**, 21 (1978).

[29] G. Margaria, E. Riccomagno, M. Chappell, and H. Wynn, *Differential algebra methods for the study of the structural identifiability of rational function state-space models in the biosciences,* Mathematical Biosciences **174**, 1 (2001).

[30] S. Vajda, *Structural identifiability of linear, bilinear, polynomial and rational systems,* IFAC Proceedings Volumes **17**, 717–722 (1984).

[31] O. Chis, J. R. Banga, and E. Balsa-Canto, *Genssi: a software toolbox for structural identifiability analysis of biological models,* Bioinformatics **27**, 2610–2611 (2011).

[32] T. Ligon, F. Fröhlich, O. Chis, J. Banga, E. Balsa-Canto, and J. Hasenauer, *Genssi 2.0: multi-experiment structural identifiability analysis of sbml models,* Bioinformatics **34**, 1421–1423 (2018).

[33] A. F. Villaverde, A. Barreiro, and A. Papachristodoulou, *Structural identifiability of dynamic systems biology models,* PLOS Computational Biology **20**, 1 (2016).

[34] H. Hong, A. Ovchinnikov, G. Pogudin, , and C. Yap, *Global Identifiability of Differential Models,* arXiv e-prints , arXiv:1801.08112 (2018).

[35] N. Meshkat, M. Eisenberg, and J. J. DiStefano, *An algorithm for finding globally identifiable parameter combinations of nonlinear ode models using Gröbner Bases,* Mathematical Biosciences **222**, 61 (2009).

[36] G. Bellu, M. P. Saccomani, S. Audoly, and L. D'Angió, *Daisy: A new software tool to test global identifiability of biological and physiological systems,* Computer Methods and Programs in Biomedicine **81**, 52 (2007).

[37] S. Audoly, G. Bellu, L. D'Angió, M. Saccomani, and C. Cobelli, *Global identifiability of nonlinear models of biological systems,* IEEE Trans Biomed Eng **48**, 55 (2001).

[38] S. Diop and M. Fliess, *Nonlinear observability,* in *In Proc. 1st Europ. Control Conf* (1991).

[39] G. Carra'Ferro and V. Gerdt, *Improved kolchin-ritt algorithm,* Program Comput Soft **29**, 83 (2003).

[40] F. Ollivier, *Le problème de l'identifiabilité globale: étude thé orique, méthodes effectives et bornes de complexité,* Ph.D. thesis, Ecole Polytechnique (1990).

[41] A. Isidori, *Nonlinear control systems* (Springer, 1995).

[42] E. D. Sontag, *Mathematical control theory: deterministic finite dimensional systems*, Vol. 6 (Springer Science and Business Media, Berlin, Germany, 2013).

[43] J. Karlsson, M. Anguelova, and M. Jirstrand, *An efficient method for structural identifiability analysis of large dynamic systems,* IFAC Proceedings Volumes **45**, 941 (2012), 16th IFAC Symposium on System Identification.

[44] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmuller, and J. Timmer, *Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood,* Bioinformatics **25**, 1923–1929 (2009).

[45] J. Jacquez and P. Greif, *Numerical parameter identifiability and estimability: Integrating identifiability, estimability, and optimal sampling design,* Mathematical Biosciences **77**, 201 (1985).

[46] D. Degenring, C. Froemel, G. Dikta, and R. Takors, *Sensitivity analysis for the reduction of complex metabolism models,* J Proc Cont **14**, 729–745 (2004).

[47] K. Gadkar, R. Gunawan, and F. Doyle, *Iterative approach to model identification of biological networks,* BMC Bioinformatics **6**, 155–175 (2005).

[48] T. Quaiser and M. Mönnigmann, *Systematic identifiability testing for unambiguous mechanistic modeling - application to jak-stat, map kinase, and nf-kb signaling pathway models,* BMC Syst Biol **3**, 50–71 (2009).

[49] L. Denis-Vidal and G. Joly-Blanchard, *An easy to check criterion for unidentifiability of uncontrolled systems and its applications,* IEEE Trans Auto Con **45**, 768–771 (2000).

[50] E. Walter, I. Braems, L. Jaulin, and M. Kieffer, *Guaranteed numerical computation as an alternative to computer algebra for testing models for identifiability,* Lecture Notes in Computer Science , 124–131 (2004).

[51] E. Walter and Y. Lecourtier, *Global approaches to identifiability testing for linear and nonlinear state space models,* Mathematics and Computers in Simulation **24**, 472 (1982).

[52] S. Vajda and H. Rabitz, *State isomorphism approach to global identifiability of nonlinear systems,* IEEE Trans Autom Cont **34**, 220–223 (1989).

[53] S. Vajda, K. Godfrey, and H. Rabitz, *Similarity transformation approach to identifiability analysis of nonlinear compartmental models,* Mathematical Biosciences **93**, 217–248 (1989a).

[54] M. Chappel and K. Godfrey, *Structural identifiability of the parameters of a nonlinear batch reactor model,* Mathematical Biosciences **108**, 245–251 (1992).

[55] X. Xia and C. Moog, *Identifiability of nonlinear systems with application to hiv/aids models,* IEEE Trans. Autom. Control **48**, 330–336 (2003).

[56] H. Wu, H. Zhu, H. Miao, and A. Perelson, *Identifiability and statistical estimation of dynamic parameters in hiv/aids dynamic models,* B Math Biol **70**, 785–799 (2008).

[57] J. D. Stigter and J. Molenaar, *A fast algorithm to assess local structural identifiability,* Automatica **58**, 118 (2015).

[58] K. Kolczyk and C. Conradi, *Challenges in horizontal model integration,* BMC Systems Biology **10** (2016), 10.1186/s12918-016-0266-3.

[59] S. Ghosh, Y. Matsuoka, Y. Asai, K. Hsin, and H. Kitano, *Software for systems biology: from tools to integrated platforms,* Nature Reviews Genetics **12**, 821–832 (2011).

[60] M. P. Saccomani, S. Audoly, and L. D'Angió, *Parameter identifiability of nonlinear systems: the role of initial conditions,* Automatica **39**, 619 (2003).

[61] M. Anguelova, J. Karlsson, and M. Jirstrand, *Minimal output sets for identifiability,* Mathematical Biosciences **239**, 139 (2012).

[62] N. Evans, S. Cheung, and J. Yates, *Structural identifiability for mathematical pharmacology: models of myelosuppression,* J Pharmacokinet Pharmacodyn **45**, 70 (2018).

[63] Y. Liu, J. Slotine, and B. A.L., *Observability of complex systems,* Proc Natl Acad Sci **10**, 2460 (2013).

[64] T. Lipniacki, P. Paszek, A. Brasier, B. Luxon, and M. Kimmel, *Mathematical model of nf-kb regulatory module,* Theor Biol **228**, 195 (2004).

[65] S. Yamada, S. Shiono, A. Joo, and A. Yoshimura, *Control mechanism of jak/stat signal transduction pathway,* FEBS Lett **534**, 190 (2003).

[66] A. Raue, V. Becker, U. Klingmüller, and J. Timmer, *Identifiability and observability analysis for experimental design in nonlinear dynamical models,* Chaos **20**, 045105 (2010).

[67] R. Heinrich and S. Schuster, *The regulation of Cellular Systems* (Berlin: Springer, 1996).

[68] A. Goldbeter, *A model for circadian oscillations in the drosophila period protein (per),* Proc Biol Sci **261**, 319 (1995).

[69] G. Margaria, E. Riccomagno, M. J. Chappell, and H. P. Wynn, *Differential algebra methods for the study of the structural identifiability of rational function state-space models in the biosciences,* Mathematical Biosciences **174**, 1 (2001).

[70] A. Holmberg and J. Ranta, *Procedures for parameter and state estimation of microbial growth process models,* Automatica **18**, 181 (1982).

[71] V. Raia, M. Schilling, M. Böhm, B. Hahn, A. Kowarsch, A. Raue, C. Sticht, S. Bohl, M. Saile, P. Möller, N. Gretz, J. Timmer, F. Theis, W.-D. Lehmann, P. Lichter, and U. Klingmüller, *Dynamic mathematical modeling of IL13-induced signaling in hodgkin and primary mediastinal B-cell lymphoma allows prediction of therapeutic targets,* Cancer Research **71**, 693 (2011).

[72] F. Bianconi, E. Baldelli, V. Ludovini, L. Crinó, A. Flacco, and P. Valigi, *Computational model of egfr and igf1r pathways in lung cancer: a systems biology approach for translational oncology,* Biotechnol Adv **Jan-Feb**, 142 (2012).

[73] F. Bianconi, E. Baldelli, V. Ludovini, L. Crinó, A. Flacco, and P. Valigi, *Egfr and Igf1r pathway in lung cancer,* (2012).

[74] A. F. Villaverde, S. Bongard, K. Mauch, D. Müller, E. Balsa-Canto, J. Schmid, and J. R. Banga, *A consensus approach for estimating the predictive accuracy of dynamic models in biology,* Computer Methods and Programs in Biomedicine **119**, 17 (2015).

[75] C. Letellier, I. Sendiña Nadal, E. Bianco-Martinez, and M. Baptista, *A symbolic network-based nonlinear theory for dynamical systems observability,* Sci Rep **8**, 1 (2018).

[76] M. P. Saccomani, S. Audoly, G. Bellu, and L. D'Angió, *Examples of testing global identifiability of biological and biomedical models with the daisy software,* Computers in Biology and Medicine **40**, 402–407 (2010).

[77] F. Mazzia, J. R.Cash, and K. Soetaert, *A test set for stiff initial value problem solvers in the open source software r: Package detestset,* Journal of Computational and Applied Mathematics **236**, 4119 (2012).

[78] Dipartimento di Matematica, University of Bari, *Test set for initial value problem solvers,* `http://www.dm.uniba.it/~testset` (2012), accessed: 2018-06-15.

[79] L. Denis-Vidal, G. Joly-Blanchard, and C. Noiret, *Some effective approaches to check the identifiability of uncontrolled nonlinear systems,* Math. Comput. Simul. **57**, 35 (2001).

[80] A. Villaverde and J. Banga, *Structural properties of dynamic systems biology models: Identifiability,reachability and initial conditions,* Processes **5** (2017), 10.3390/pr5-20029.

[81] J. Graciano, D. Mendoza, and G. L. Roux, *Performance comparison of parameter estimation techniques for unidentifiable models,* Computers and Chemical Engineering **64**, 24–40 (2014).

1

APPENDIX

Table 1.1: **Summary of the models analysed in this thesis.**

Name	States	System Parameters	Total unknowns	Chapter
A chemical reaction system [63]	11	6	17	2
NF-κB model [64]	15	28	43	2
JAK/STAT model [61, 65]	31	51	82	2, 4
Ligand binding model [66]	6	8	14	2
Simplified glycolytic reaction model [67]	10	13	23	2
Goldbeter model [68]	5	17	22	2
Reparameterised JAK/STAT model with specific model output	14	20	21	2
Irrational JAK/STAT model with specific model output	14	20	21	2
Immunological model for mastitis in dairy cows [69]	2	5	7	3
Microbial growth model [70]	2	4	6	3
JAK/STAT model [71]	14	22	23	3, 4, 5
Lung cancer model [72, 73]	21	54	75	3, 4
Chinese Hamster model [74]	34	117	151	4
Novak Tyson model [75]	13	39	52	4
Model with 20 states [76]	20	22	42	4
Pollution model [77, 78]	20	25	45	4
Small benchmark model [79]	2	3	3	5
Model with multiple sets of problematic initial conditions	2	1	3	5
Benchmark model with input [60]	2	4	4	5
Model describing a simple biochemical network [80]	3	3	6	5
Three-phase industrial batch reactor [81]	7	5	5	5

2

DETERMINING MINIMAL OUTPUT SETS THAT ENSURE STRUCTURAL IDENTIFIABILITY

Dominique JOUBERT, Hans STIGTER, Jaap MOLENAAR

The minimal output set problem has not received much attention previously, possibly due to the difficulty of testing structural identifiability in the first place.

(Anguevola, Karlsson, Jirstrand, 2012)

Published as PLoS ONE **13**(11):e0207334. (2018) [1].

ABSTRACT

THE process of inferring parameter values from experimental data can be a cumbersome task. In addition, the collection of experimental data can be time consuming and costly. This paper covers both these issues by addressing the following question: "Which experimental outputs should be measured to ensure that unique model parameters can be calculated?". Stated formally, we examine the topic of minimal output sets that guarantee a model's structural identifiability. To that end, we introduce an algorithm that guides a researcher as to which model outputs to measure. Our algorithm consists of an iterative structural identifiability analysis and can determine multiple minimal output sets of a model. This choice in different output sets offers researchers flexibility during experimental design. Our method can determine minimal output sets of large differential equation models within short computational times.

2.1. INTRODUCTION

Mathematical models are powerful tools that enable the scientific community to understand processes otherwise immeasurable by predicting outcomes of numerous physical properties. The field of systems biology often utilises ordinary differential equations to model dynamic systems. These models can comprise large systems of differential equations that contain vast numbers of unknown parameters [2]. Despite improvements in the quality of experimental sensors and therefore both the quality and quantity of experimental data, the process of parameter estimation remains cumbersome. This may be due to noisy data or due to the inherent structure of the model (structural unidentifiability) [3]. A structurally unidentifiable model implies that certain parameters are totally correlated, also referred to as 'aliased', and have confidence intervals that span the interval $(-\infty, \infty)$. Uncertainty in inferred parameter values calls into question the validity of the entire model and therefore it is imperative to address these uncertainties upfront by conducting identifiability analyses.

We will focus on ensuring structural identifiability and since this property can be analysed before conducting experiments, our analysis can be utilised in preliminary experimental design. An experimental researcher may wish to know: "Which of the predefined model outputs/sensors do I at least need to measure to ensure that I can infer unique parameter values?". The answer is addressed by the topic of minimal output sets, where a minimal output set is defined as: *Measuring a minimal set of model outputs ensures that a model is structurally identifiable.* Due to its complexity, the topic of minimal output sets has received little attention [4]. Scientists often rely on intuitive experimental design, which may easily result in redundant or insufficient experimental measurements.

In this paper we present an algorithm to determine minimal output sets by identifying sets of totally correlated parameters using an iterative structural identifiability analysis. This algorithm offers insight into which sensors should be measured, thereby aiding intuitive experimental design. A particular model may have multiple minimal output sets. This offers great flexibility to the experimental researcher as he/she can decide which output set to measure taking factors such as time, cost and physical constraints into account.

This structural identifiability issue has been considered in a few previous papers [4–7], which we will briefly describe. The first paper, published in 2009, introduces a minimisation algorithm to determine which parameters are identifiable [5]. Three simple examples are included and due to its computational complexity the author states that defining minimal output sets for medium sized models is still too hard using this algorithm. In a paper published in 2012, the authors present an algorithm tasked with identifying symmetries, i.e. sets of totally correlated parameters, in a system of differential equations [4]. Once these symmetries have been identified, the states and parameters that destroy these symmetries are included into minimal output sets. Minimal output sets of the well-known NF-κB and JAK/STAT models are determined assuming that all model parameters and states can potentially be measured. The final step in their algorithm is doing a symbolic computation to test for structural identifiability and identify any remaining symmetries. Other papers address observability [6, 7]. Identifiability can be regarded as a special case of observability [8]. In [6], the authors introduce a graph-

ical method and illustrate its key concepts using nonlinear models. They construct a directed graph from the so-called adjacency matrix and inspect it to identify strongly connected components and more specifically root strongly connected components. A directed graph is a graphical representation of an ODE system and depicts the connectivity between individual states. Two nodes are classified as strongly connected if they are reachable from each other [9]. Root strongly connected components are strongly connected components with no outgoing edges. Minimum output sets are identified from the different elements in these root strongly connected components. A different approach is followed by Letellier and co-authors [7]. They use a symbolically computed Jacobi matrix to compute the output sets that ensure observability. An interesting extension of minimal output sets in the preliminary experimental design phase, could be to determine these sets taking measurement noise into account, thereby establishing practical identifiability. To this end, Docherty and co-authors present a graphical method to identify such sets [10].

Our minimal output set algorithm is different from the existing techniques as it numerically identifies sets of unidentifiable parameters. Through a number of computational experiments, we provide evidence (but not a complete mathematical proof) that our proposed algorithm has the following attributes:

- It can calculate the minimal output sets of large models.
- It can easily be adjusted to allow for cases in which only a limited subset of predefined outputs are measurable. This is illustrated in example 7 in the results and discussion section.
- Irrational models can also be analysed as shown in example 8 in the results and discussion section.

The numerical findings are validated in a second step using symbolic computations as explained in [11]. This paper is divided into the following sections: Section 2.2 covers the underlying theory and concepts of our algorithm. Section 2.3 showcases the algorithm using 8 examples and the final section contains concluding remarks.

2.2. MATERIALS AND METHODS

BACKGROUND THEORY

Many dynamic systems biology phenomena are described in terms of differential equation models. These models can often be written in the standard state-space form [12]:

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{f}(t, \boldsymbol{x}(t), \boldsymbol{\theta}), \tag{2.1}$$

$$\boldsymbol{x}(0) = \boldsymbol{x_0}, \tag{2.2}$$

$$\boldsymbol{y}(t) = \boldsymbol{h}(\boldsymbol{x}(t), \boldsymbol{\theta}). \tag{2.3}$$

State variables are contained in a vector $\boldsymbol{x}(t)$ (dim $\boldsymbol{x} = n$), model parameters are contained in vector $\boldsymbol{\theta}$, (dim $\boldsymbol{\theta} = p$) and the output signals or measured variables are contained in vector $\boldsymbol{y}(t)$ (dim $\boldsymbol{y} = m$). Function $\boldsymbol{f}$ denotes a dynamic model structure and $\boldsymbol{h}$ is the output or observation function. Our approach allows for functions $\boldsymbol{f}$ and $\boldsymbol{h}$ to be

either rational or irrational. Unknown initial conditions of model states in vector $\boldsymbol{x}_0$, can be regarded as additional unknown parameters and can accordingly be included into $\boldsymbol{\theta}$. If all initial conditions are unknown, $\boldsymbol{\theta}$ contains $p+n$ elements.

The identifiability analysis method used in this paper was first proposed in [11]. In essence, this method relies on the singular value decomposition (SVD) of an output sensitivity matrix. Reid introduced the concept of sensitivity based identifiability analysis for linear models [13]. In his paper, he defines a sensitivity matrix as $\boldsymbol{S} = \partial \boldsymbol{y}/\partial \boldsymbol{\theta}$, with its elements describing the sensitivities of the model output with respect to model parameters. These partial derivatives are evaluated for nominal parameter values $\boldsymbol{\theta_0}$. Let $\Delta\boldsymbol{\theta}$ denote a small perturbation of the nominal vector $\boldsymbol{\theta_0}$, so $\boldsymbol{\theta} = \boldsymbol{\theta_0} + \Delta\boldsymbol{\theta}$. This perturbation will result in a corresponding perturbation in the model output, $\boldsymbol{y}(\boldsymbol{\theta}) = \boldsymbol{y}(\boldsymbol{\theta_0}) + \Delta\boldsymbol{y}$. A first order Taylor series approximation can be used to relate these perturbations [14, 15]:

$$\boldsymbol{y}(\boldsymbol{\theta}) - \boldsymbol{y}(\boldsymbol{\theta_0}) \approx \boldsymbol{S} \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \quad \text{or} \quad \Delta\boldsymbol{y} \approx \boldsymbol{S} \cdot \Delta\boldsymbol{\theta}. \tag{2.4}$$

To solve $\Delta\boldsymbol{\theta}$ from the measured $\Delta\boldsymbol{y}$ uniquely, $\boldsymbol{S}^T\boldsymbol{S}$ should be nonsingular [16–18] and therefore $\boldsymbol{S}$ should be of full rank [19, 20]. For nonlinear models, the individual sensitivities are obtained by deriving the model (2.1)-(2.3) with respect to $\boldsymbol{\theta}$, thereby obtaining the system:

$$\frac{d}{dt}\left(\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\theta}}\right) = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\theta}} + \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\theta}}, \tag{2.5}$$

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{\theta}} = \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}}\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\theta}} + \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{\theta}}. \tag{2.6}$$

To obtain the output sensitivity matrix $\boldsymbol{S}$, the matrix function $\partial \boldsymbol{y}/\partial \boldsymbol{\theta}$ is evaluated over a discretised finite time grid, $[t_0, \ldots, t_N]$, and the obtained matrices at each time point are vertically concatenated [11]. It is advantageous to normalise $\boldsymbol{S}$ to adjust for sensitivities measured in different units [21]. We emphasise that working with normalised matrix elements might be numerically attractive but not essential. The normalised matrix $\boldsymbol{S}_{norm}$ is given as:

$$\boldsymbol{S}_{norm} = \begin{pmatrix} \frac{\theta_1}{y_1(t_0)}\frac{\partial y_1}{\partial \theta_1}(t_0) & \cdots & \frac{\theta_{p+n}}{y_1(t_0)}\frac{\partial y_1}{\partial \theta_{p+n}}(t_0) \\ \vdots & \ddots & \vdots \\ \frac{\theta_1}{y_m(t_0)}\frac{\partial y_m}{\partial \theta_1}(t_0) & \cdots & \frac{\theta_{p+n}}{y_m(t_0)}\frac{\partial y_m}{\partial \theta_{p+n}}(t_0) \\ \vdots & & \vdots \\ \frac{\theta_1}{y_1(t_N)}\frac{\partial y_1}{\partial \theta_1}(t_N) & \cdots & \frac{\theta_{p+n}}{y_1(t_N)}\frac{\partial y_1}{\partial \theta_{p+n}}(t_N) \\ \vdots & \ddots & \vdots \\ \frac{\theta_1}{y_m(t_N)}\frac{\partial y_m}{\partial \theta_1}(t_N) & \cdots & \frac{\theta_{p+n}}{y_m(t_N)}\frac{\partial y_m}{\partial \theta_{p+n}}(t_N) \end{pmatrix}. \tag{2.7}$$

If all the initial conditions of model states are unknown, matrix $\boldsymbol{S}_{norm}$ (and also $\boldsymbol{S}$) has dimensions $M\times(p+n)$, with $M = m\times(N+1)$. To determine the rank of $\boldsymbol{S}_{norm}$ (and also $\boldsymbol{S}$), the numerical rank test using an SVD reads as [16]:

$$\boldsymbol{S}_{norm} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T. \tag{2.8}$$

The 2 matrices of importance are, the diagonal matrix, $\boldsymbol{\Sigma}$ (dim $M\times(p+n)$), and $\boldsymbol{V}$ (dim $(p+n)\times(p+n)$). The singular values in $\boldsymbol{\Sigma}$, $\sigma_i, i = 1,\ldots,p+n$, are used to determine whether or not $\boldsymbol{S}_{norm}$ (or $\boldsymbol{S}$) is of full rank. The rank of $\boldsymbol{S}_{norm}$ (or $\boldsymbol{S}$) is the number of nonzero singular values and this can be expressed as follows [16]:

$$\text{If } \sigma_1 \geq \ldots \geq \sigma_q > \sigma_{q+1} = \ldots = \sigma_{p+n} = 0, \text{then rank } (\boldsymbol{S}_{norm}) = q. \tag{2.9}$$

In practice, singular values are never exactly vanishing due to numerical rounding errors. That is why one uses as practical definition: zero-valued singular values are values that fall beyond a significant gap in the spectrum of singular values [22]. In this paper we consider a gap larger than 3 decades on the log scale as significant. Once structural unidentifiability has been established, the nonzero entries of the singular vectors of matrix $\boldsymbol{V}$, related to vanishing singular values beyond this gap, allude to which model parameters and initial conditions may be unidentifiable. The singular values and the unidentifiable parameters are graphically illustrated in a so-called identifiability signature [23].

To illustrate our approach, we use the NF-κB model, also analysed in Section 3. It has 15 states and 28 model parameters and if all the initial conditions of the individual model states are considered to be unknown, it has a total of 43 parameters [4]. Measuring $\boldsymbol{y_{max}} = [x_1,\ldots,x_{15}]$ as model output, we observe no gap in the singular values (see Fig 2.1). This confirms that there are no vanishing singular values and therefore the sensitivity matrix, $\boldsymbol{S}_{norm}$, is of full rank and the model is structurally identifiable for this particular choice of output sensors.

However, if we omit state x_4 from the output $\boldsymbol{y_{max}}$, we observe from Fig 2.2 that matrix $\boldsymbol{S}_{norm}$ is now rank deficient. This is apparent from the clear gap in the singular values and the vanishing singular value of $\sigma_{43} = 7.8\times10^{-16}$.

We can now examine the columns of $\boldsymbol{V}$, corresponding to vanishing singular values, for suggestions as to which model parameters may be unidentifiable. Fig 2.2 reveals only 1 vanishing singular value and therefore it suffices to consider only the last column vector, $\boldsymbol{v}_{43}$, corresponding to σ_{43}. The nonzero entries in Fig 2.3 reveal that parameters $\theta_2,\theta_3,\theta_{27}$ and the initial condition $x_4(0)$, are both totally correlated and unidentifiable. To ensure the model's structural identifiability, the omitted state, x_4, has to be measured and so is included into any minimal output set. In contrast, omitting state x_3 from the output set does not change this model's identifiability and therefore can be omitted from a minimal output set.

MINIMAL OUTPUT SET ALGORITHM

Here, we present our algorithm to detect minimal output sets. We first outline the ideas underlying the algorithm and then discuss the subsequent steps. It is important to realise that the parameters to be identified may comprise both system parameters $\theta_j, j =$

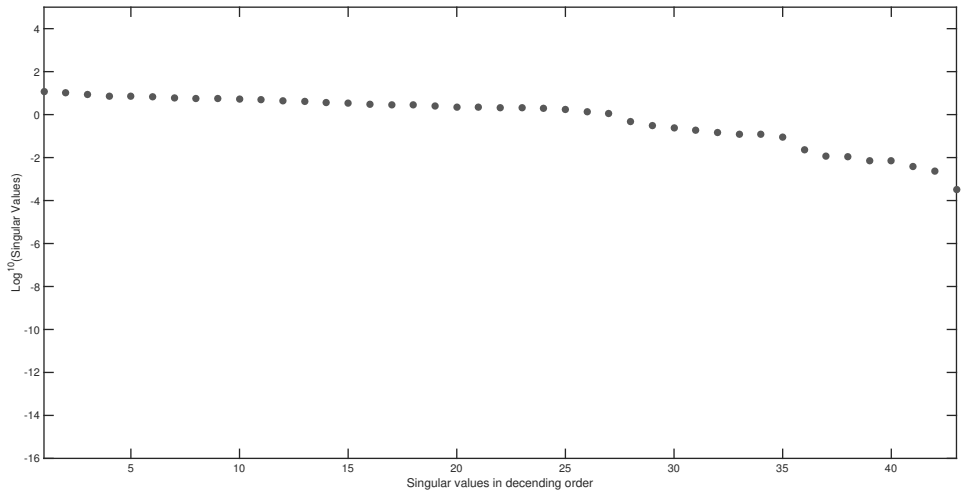


Figure 2.1: **NF-κB model: Singular values of the output sensitivity matrix, S_{norm}, if we measure all states, $\{x_1,\ldots,x_{15}\}$, as model output.** Singular values, arranged in descending order, reveal no gap. This suggests that the sensitivity matrix is of full rank and therefore the model is structurally identifiable.

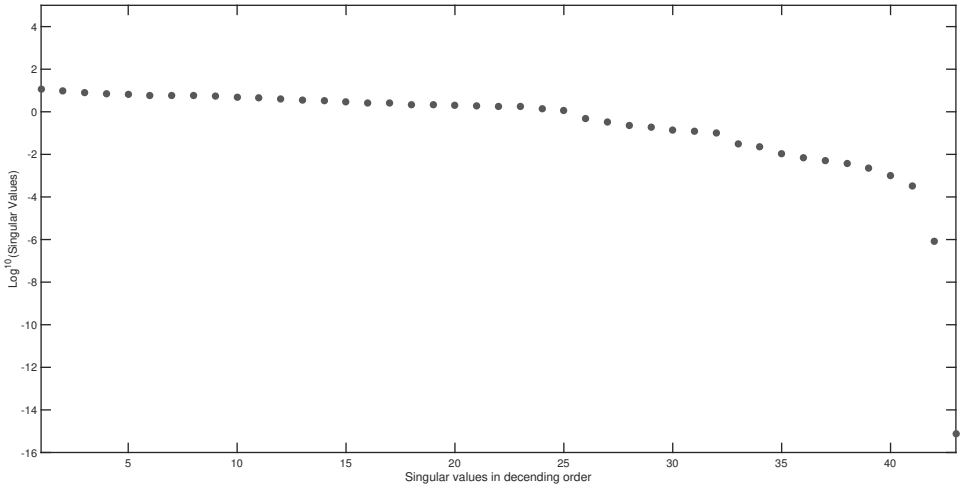


Figure 2.2: **NF-κB model: Singular values of the output sensitivity matrix, S_{norm}, if we measure all states apart from x_4.** Singular values, arranged in descending order reveal a clear gap with $\sigma_{43} = 7.8\times10^{-16}$. This indicates that the sensitivity matrix is rank deficient and so the model is structurally unidentifiable.

1,..,p, and initial values of the states, $x_j(0), j = 1,\ldots,n$. We assume that the numerical values assigned to the elements in both $\boldsymbol{\theta}$ and $\boldsymbol{x}(0)$ are regular points, where it is known that the rank of the sensitivity matrix does not change in the neighbourhood of a regular point. To ensure that this assumption holds, it may be useful to repeat the algorithm for a different values in the vicinity of a chosen regular point. Furthermore, system pa-

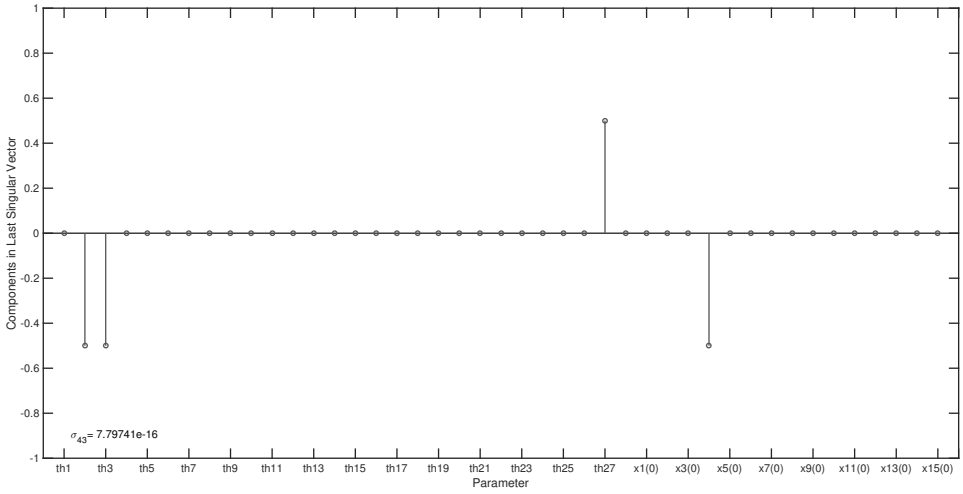


Figure 2.3: **NF-κB model: Entries in the last right singular vector corresponding to the vanishing singular value, σ_{43}, in Fig 2.2.** The corresponding nontrivial null-space indicates that parameters $\theta_2, \theta_3, \theta_{27}$ and initial condition $x_4(0)$ are totally correlated.

rameters are to be *inferred* from measurements of model outputs and so are usually not regarded as measurable outputs. For the time being, we assume that the predefined measurable outputs $y_j, j = 1,\ldots,m$, also referred to as sensors, are identical to the states $x_j, j = 1,\ldots,n$, and therefore $m = n$. Later on we show that this assumption can easily be relaxed. We may also take for granted that the system is identifiable when all sensors are measured. If this would not be the case, searching for minimal output sets would clearly not be possible.

The main idea of the algorithm is to systematically omit elements from the set of all available states/sensors, thereby searching for essential sensors that absolutely can not be omitted to keep the system identifiable. As explained above, unidentifiability is detected by inspecting the calculated singular vales of the sensitivity matrix in (2.7). If these singular values show a gap of 3 decades or larger, we conclude unidentifiability and subsequently proceed to identify the essential states/sensors that need to be included into a model's minimal output sets.

Let $\boldsymbol{y}_{max}$ be the set of all available states/sensors with set-cardinality $|\boldsymbol{y}_{max}| = m$. The algorithm involves an iterative identifiability analyses in which sensors are omitted step-wise from the maximum starting set $\boldsymbol{y}_{max}$. Systematically more and more sensors are left out as to find all essential sensors that are needed for a minimal output set (MOS).

Let k be the number of sensors to be omitted from a set of available sensors, $\boldsymbol{y}^k$. Starting with $k = 1$, we leave out one-sensor-at-a-time from the initial set of *all available sensors*, $\boldsymbol{y}^1 := \boldsymbol{y}_{max}$. Each time measuring with a different set of sensors from $\boldsymbol{y}^1$, we conduct $\binom{m}{m-1} = m$ identifiability analyses for $k = 1$. If a lack of identifiability is detected, the unidentifiable parameters are stored in a set ϕ_1 and the corresponding omitted sensors that cause unidentifiability are stored in a set ψ_1. Continuing this way, we get unidentifiable parameter sets $\{\phi_i, i = 1,\ldots,l_1\}$, and the corresponding omitted sen-

sor sets $\{\psi_i, i = 1,\ldots,l_1\}$ that cause a lack of identifiability. Here, l_1 is the total number of unidentifiable parameter sets identified for the case of omitting one-sensor-at-a-time ($k = 1$). The unidentifiable parameter sets ϕ_i can be found by inspecting the nonzero entries in the singular vectors of the matrix $\boldsymbol{V}$ corresponding with the zero-valued singular values (as can be seen from the identifiability signature).

To ensure structural identifiability, the essential sensor form each $\{\psi_i, i = 1,\ldots,l_1\}$ *must be* included into *any* minimal output set. Having checked all possibilities of leaving out one-sensor-at-a-time, we can now define a new set of available sensors, say $\boldsymbol{y}^2$, that is created by excluding the previously found sensors in the sets $\{\psi_i, i = 1,\ldots,l_1\}$ from the set $\boldsymbol{y}^1$. Since we know for sure that these excluded sensors are needed for a model's structural identifiability, they are permanently included into all sensor sets that are measured from now on. Hence, the case $k = 1$ *reduces* the number of candidate sensors to choose from in the next iteration from m to $m' = m - l_1$.

Next, we leave out two-sensors-at-a-time (the case $k = 2$) from $\boldsymbol{y}^2$ and check for identifiability. Since set cardinality $|\boldsymbol{y}_2|$ now equals $m' \leq m$, we have $\binom{m'}{m'-2}$ choices for omitting 2 sensors from this set. If unidentifiability is detected, a new set of unidentifiable parameters is compiled from the identifiability signature and stored in ϕ_{l_1+1}, and the 2 corresponding left-out sensors are stored in ψ_{l_1+1}. Proceeding this way, the total number of unidentifiable sets that can be found for $k = 2$ are collected in the sets $\{\phi_i, i = l_1+1,\ldots,l_1+l_2\}$ and the corresponding omitted sensor sets, $\{\psi_i, i = l_1+1,\ldots,l_1+l_2\}$.

Assume now that for the case of leaving out two-sensors-at-a-time ($k = 2$), we have found an unidentifiable parameter set ϕ_i. Apparently, this new set ϕ_i, *only occurs when 2 particular sensors, recorded in the corresponding set* ψ_i, *are missing* and therefore either 1 of these 2 essential sensors *must be* included in a MOS. Hence, the available sensor sets for the case $k = 3$ branch out into two sets, namely $\boldsymbol{y}^{3,1}$ and $\boldsymbol{y}^{3,2}$. When leaving out three-sensors-at-a-time in the next iteration of our algorithm (case $k = 3$), we have to iterate both of these available sensor sets to find more unidentifiable parameter sets $\{\phi_i, i = l_1+l_2+1,\ldots,l_1+l_2+l_3\}$. Continuing in this way for $k = 3,4,\ldots$, we complete our search for essential sensors when leaving out k sensors at a time. At the same time guaranteeing that the sensors that are needed for the identifiability of a model, identified in earlier iterations $k-1, k-2,\ldots,1$, are included in each new measured output.

Clearly, for large models the output, $\boldsymbol{y}_{max}$, will contain a large number of sensors and in these cases an exhaustive search will be computationally demanding. The computational burden may however be substantially reduced by randomly selecting outputs from an intermediate set of available sensors $\boldsymbol{y}^k$ (for a certain iteration step k) using a series of Bernoulli trial experiments. The number of sensors to include into each sensor set can then be chosen in such a way that the chance of successfully detecting an unidentifiable set of parameters is more than 99.5% (refer to supplementary S9 File in the appendix).

We further note that in practice our experience shows that the values $k = 1,2,3$ already summarise the *majority* of possible unidentifiable parameter sets ϕ_i. More importantly, once we have established a few required sensors on basis of lower k values, one can perform an additional check for a lack of identifiability when *using only the required sensors that have already been determined for the lower k values.* Such a test will immediately reveal additional correlations that still need to be found for larger k values,

but these correlations are not yet neatly separated in a systematic way. This check does, however, demonstrate whether we need to continue our search for larger k values (e.g. $k = 4,5,\ldots$), yes or no or whether one can already define minimal outputs sets from the already identified essential sensors.

Finally, in reality the output, $\boldsymbol{y}_{max}$, is not always identical to the states $\boldsymbol{x}$. For example, one could have $\boldsymbol{y}_{max} = [x_1 + x_3 + x_4, \theta_{16}(x_3 + x_4 + x_5 + x_{12}), \theta_{17}(x_4 + x_5)]$. Our algorithm allows for the user to define these more complex outputs in a straightforward manner: Instead of omitting states $\{x_i, i = 1,\ldots,n\}$, we now systematically omit *outputs* $\{y_j, j = 1,\ldots,m\}$ to find the essential sensors needed in a MOS.

2.3. RESULTS AND DISCUSSION

EXAMPLE 1: A CHEMICAL REACTION SYSTEM

This model was used by Liu and co-authors to illustrate their method ensuring observability based on the graphical analysis of a model's structure [6]. It contains 11 states and 6 model parameters and potentially has 17 unknown parameters. Examining the structure of the model by evaluating its adjacency/Jacobi matrix, the authors detected 3 root strongly connected components and identified 6 minimal output sets.

These observability results were confirmed using our algorithm. Additionally, we expanded the scope of the problem to define minimal output sets that guarantee this model's structural identifiability. We found that the minimal output sets that ensure observability also ensure identifiability and these are: $\{x_4, x_6, x_7\}$, $\{x_4, x_6, x_8\}$, $\{x_4, x_6, x_9\}$, $\{x_5, x_6, x_7\}$, $\{x_5, x_6, x_8\}$ and $\{x_5, x_6, x_9\}$. These results were obtained in 6 minutes and 35 seconds using a Intel Core i7 processor with 8GB RAM (see S1 File in the appendix for details).

Using our algorithm, we detected 3 different sets of unidentifiable parameters, $\{\phi_1, \phi_2, \phi_3\}$ (see S1 File for a graphical illustration of the branching of this analysis). Each of these sets can be verified symbolically, which also allows for the identification of different totally correlated sets of parameters within each set, ϕ_i (see supplementary S8 file in the appendix for the symbolic verification of all 3 unidentifiable sets). The results obtained for the different values of k are summarised in Table 2.1.

Figs 2.4 and 2.5 indicate the identifiability signature obtained when measuring the output, $\{x_1, x_2, x_3, x_6, x_7, x_8, x_9, x_{10}, x_{11}\}$, here $k = 2$. The 4 zero-valued singular values indicate that the model is unidentifiable when measuring this output. The unidentifiable parameters can be identified by looking at the nonzero entries in the last 4 columns of matrix $\boldsymbol{V}$, each corresponding to a singular value beyond the gap. Fig 2.5 reveals the unidentifiable parameter set, $\phi_2 = \{\theta_2, \theta_3, x_4(0), x_5(0)\}$ and accordingly, the essential sensors are $\psi_2 = \{x_4, x_5\}$. The symbolic verification of this set yields a nontrivial null-space with 4 base vectors: $\mathcal{N}\left(\frac{d\boldsymbol{G}}{d\boldsymbol{\theta}^{\boldsymbol{unid}}}(\boldsymbol{\theta})\right) = \{1,0,0,0\}, \{0,1,0,0\}, \{0,0,1,0\}, \{0,0,0,1\}$, where $\boldsymbol{\theta}^{unid} = \{\theta_2, \theta_3, x_4(0), x_5(0)\}$.

EXAMPLE 2: NF-κB MODEL

This model describes the two-feedback-loop regulatory module of nuclear factor NF-κB signalling pathway. It involves two-compartment kinetics of the activators IκB (IKK) and NF-κB, the inhibitors, A20 and IκBα, and their complexes. In response to extra-cellular

Table 2.1: **Results obtained during an exhaustive analysis of the chemical reaction model.**

k	Number of sets	Unidentifiable parameters sets	Omitted sensors	Computational time (sec)
1	1	$\phi_1 = \{x_6(0)\}$	$\psi_1 = \{x_6\}$	4.5
2	1	$\phi_2 = \{\theta_2, \theta_3, x_4(0), x_5(0)\}$	$\psi_2 = \{x_4, x_5\}$	14.6
3	1	$\phi_3 = \{\theta_4, \theta_5, x_7(0), x_8(0), x_9(0)\}$	$\psi_3 = \{x_7, x_8, x_9\}$	53.6
4	0			53.5
5	0			112.3
6	0			90.4
7	0			48.7
8	0			17.1

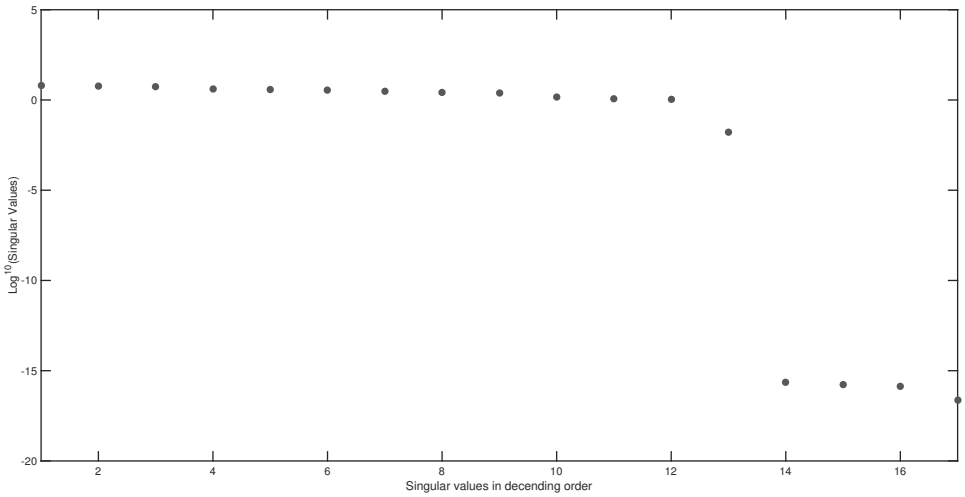


Figure 2.4: **Example 1: Structural identifiability results of a chemical reaction system: Singular values of the output sensitivity matrix, S_{norm}, when measuring the output $\{x_1 \ldots, x_{11}\}$ omitting sensors x_4 and x_5.** Singular values, arranged in descending order, reveal a clear gap. This gap, in conjunction with the smallest singular value, $\sigma_{17} = 2.4 \times 10^{-17}$, indicate that the model is structurally unidentifiable when measuring this output.

signals such as tumour necrosis factor, the activation of IKK ultimately stimulates the release of the main activator NF-κB, which enters the nucleus and triggers transcription of the inhibitors and numerous other genes [24] (See supplementary S2 File in the appendix for a model description). The model contains 15 states and 28 model parameters and assuming the initial state conditions to be unknown, it has 43 unknown parameters in total.

A minimal output set for this model was first identified by Anguelova and co-authors [4]. We found the model structural identifiable when measuring all states, $\boldsymbol{y_{max}} = \{x_1, \ldots, x_{15}\}$. Our algorithm identified 5 different sets of unidentifiable parameters: $\phi_1 = \{\theta_2, \theta_3, \theta_{27}, x_4(0)\}$, $\phi_2 = \{\theta_5, \theta_6, \theta_{18}, x_5(0)\}$, $\phi_3 = \{\theta_8, \theta_9, \theta_{10}, x_6(0)\}$, $\phi_4 = \{\theta_{19}, \theta_{27}, x_{10}(0)\}$ and $\phi_5 =$

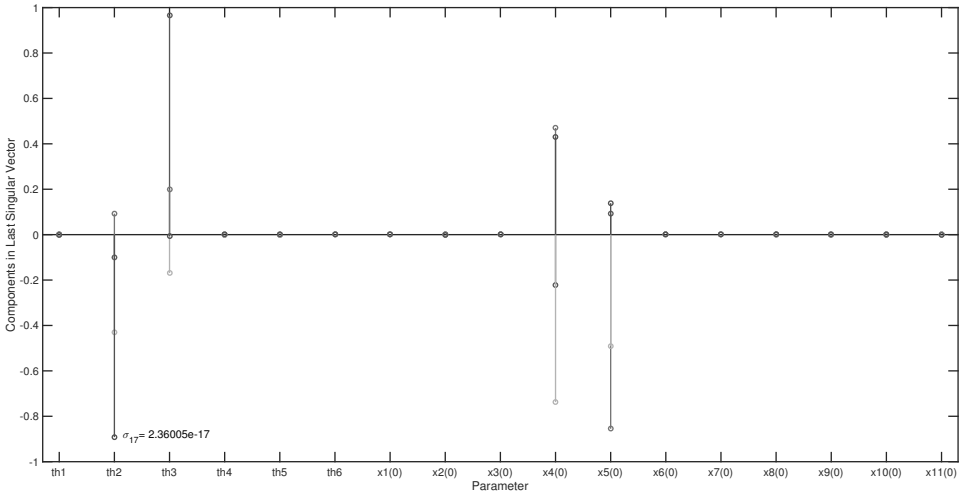


Figure 2.5: **Example 1: Structural identifiability results of a chemical reaction system: Nonzero entries in the last 4 columns of matrix V.** These indicate that initial conditions $x_4(0)$ and $x_5(0)$ and model parameters θ_2 and θ_3 are unidentifiable. Since x_4 and x_5 are simultaneously omitted from $\boldsymbol{y_{max}}$, both of these sensors are essential.

$\{x_{12}(0)\}$. The corresponding sets of essential sensors are: $\psi_1 = \{x_4\}$, $\psi_2 = \{x_5\}$, $\psi_3 = \{x_6\}$, $\psi_4 = \{x_{10}\}$ and $\psi_5 = \{x_{12}\}$ and these results were obtained in 29.5 seconds. Analysing the model for all the different values of k took 8 minutes and 20 seconds. The resulting minimal output set, $\{x_4, x_5, x_6, x_{10}, x_{12}\}$, is identical to the minimal output set defined by Anguelova *et. al.* [4].

Figs 2.2 and 2.3 show the identifiability signature obtained when sensor x_4 is omitted from $\boldsymbol{y_{max}}$. The symbolic verification of the unidentifiable set shown in Fig 2.3 yields the nontrivial null-space: $\mathcal{N}\left(\frac{d\boldsymbol{G}}{d\boldsymbol{\theta}^{\boldsymbol{unid}}}(\boldsymbol{\theta})\right) = \{\theta_2/x_4(0), -\theta_3/x_4(0), -\theta_{27}/x_4(0), 1\}$, were $\boldsymbol{\theta}^{unid} = \{\theta_2, \theta_3, \theta_{27}, x_4(0)\}$. Refer to the supplementary S8 File for symbolic verification of the remaining 4 sets of unidentifiable parameters.

EXAMPLE 3: JAK/STAT MODEL

This model aims to describe the interaction of the suppressor cytokine signaling-1 (SOCS1), Janus kinase (JAK) and the transcription (STAT) signal transduction pathway [25] (S3 File in the appendix). It contains 31 model states and 51 model parameters and so the potential total number of unknown parameters is 82. This model was structurally identifiable when measuring all states, $\boldsymbol{y_{max}} = \{x_1, \ldots, x_{31}\}$. Applying our method, we identified 2 sets of unidentifiable parameters: $\phi_1 = \{x_{31}(0)\}$ with corresponding omitted sensor set $\psi_1 = \{x_{31}\}$, and $\phi_2 = \{\theta_{14}, \theta_{51}, x_{10}(0), x_{11}(0)\}$ with corresponding omitted sensor set $\psi_2 = \{x_{10}, x_{11}\}$. These results were obtained in 3 minutes and 2 seconds with sets ψ_1 and ψ_2 identified using iterative Bernoulli trails. Measuring either x_{10} or x_{11}, the 2 identified minimal output sets of the JAK/STAT model $\{x_{10}, x_{31}\}$ or $\{x_{11}, x_{31}\}$, are identical to the findings of Anguelova and co-authors [4].

The identifiability signature obtained when states x_{10} and x_{11} are simultaneously omitted from the model's output is illustrated in Figs 2.6 and 2.7. The unidentifiable set illustrated in Fig 2.7, was confirmed by the symbolically computed nontrivial null-space: $\mathcal{N}\left(\frac{d\boldsymbol{G}}{d\boldsymbol{\theta}^{\boldsymbol{unid}}}(\boldsymbol{\theta})\right) = \{\theta_{14}/x_{11}(0), -\theta_{51}/x_{11}(0), x_{10}/x_{11}(0), 1\}$. Here $\boldsymbol{\theta}^{unid}$ is set ϕ_2.

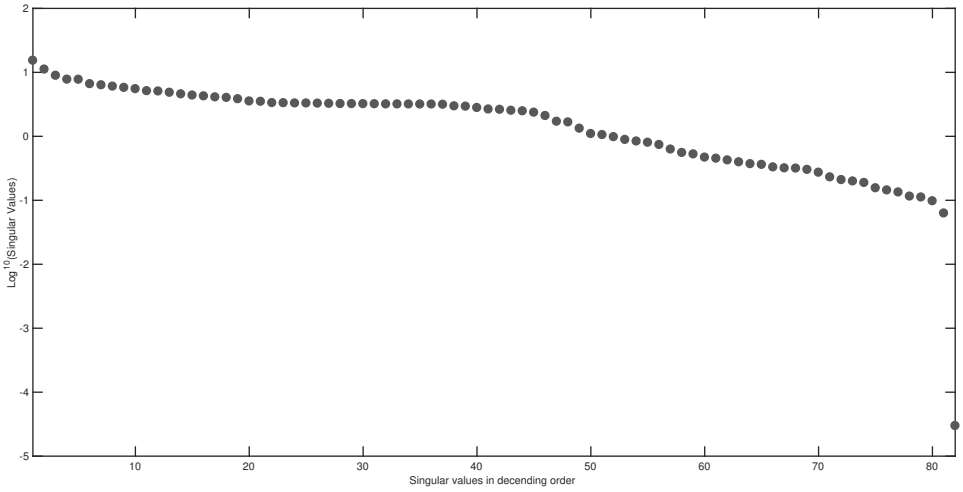


Figure 2.6: **Example 3: JAK/STAT model: Singular values of the output sensitivity matrix, $\boldsymbol{S_{norm}}$, when measuring the model output $\{\boldsymbol{x_1},,\ldots,\boldsymbol{x_{31}}\}$, omitting sensors $\boldsymbol{x_{10}}$ and $\boldsymbol{x_{11}}$.** Singular values reveal a clear gap and this, in conjunction with the smallest singular value of $\sigma_{82} = 7.3\times10^{-16}$, indicates that $\boldsymbol{S}_{norm}$ is not of full rank and therefore the model is structurally unidentifiable.

EXAMPLE 4: LIGAND BINDING MODEL

Next, we consider a Ligand binding model, previously analysed for structural identifiability [26]. This model describes the dynamic behaviour of the ligand (Epo) and its receptor (EpoR) in erythroid progenitor cells. In these cells, the dynamic characteristics of the Epo receptor (EpoR) determine how signals are encoded, in the presence of Epo, and processed at receptor level. These processed signals activate downstream signalling cascades such as the JAK2-STAT5 pathway which in turn lead to responses such as differentiation and proliferation of erythrocytes [26]. The model consists of 6 states and assuming their initial states are unknown, it contains 14 unknown parameters (see supplementary S4 File in the appendix).

The minimal output set ensuring the observability of this model, $\{x_5, x_6\}$, was determined by Liu and co-authors using their graphical approach [6]. This set also ensures the structural identifiability of the model and this result was obtained in 12 seconds. Two sets of unidentifiable parameters were detected: $\phi_1 = \{x_5(0)\}$ and $\phi_2 = \{x_6(0)\}$. Set ϕ_2, shown in Fig 2.8, is indicated by the nonzero entry in the last right singular vector corresponding to the smallest singular value calculated to be precisely zero.

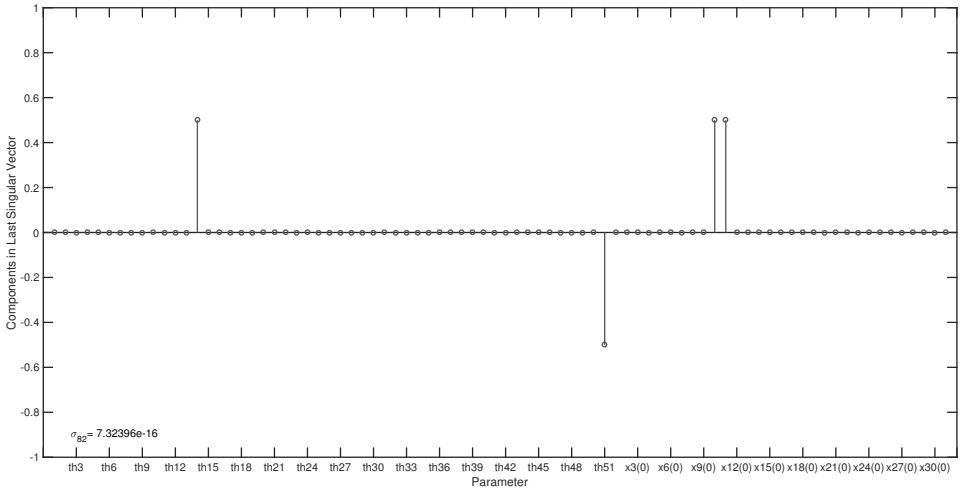


Figure 2.7: **Example 3: JAK/STAT model: Entries in the last right singular vector corresponding to the vanishing singular value, σ_{82}, in Fig 2.6.** The corresponding nontrivial null-space indicates that system parameters θ_{14}, θ_{51} and initial conditions $x_{10}(0)$ and $x_{11}(0)$ are totally correlated and so the model is not identifiable when model states x_{10} and x_{11} are simultaneously omitted from the model's output.

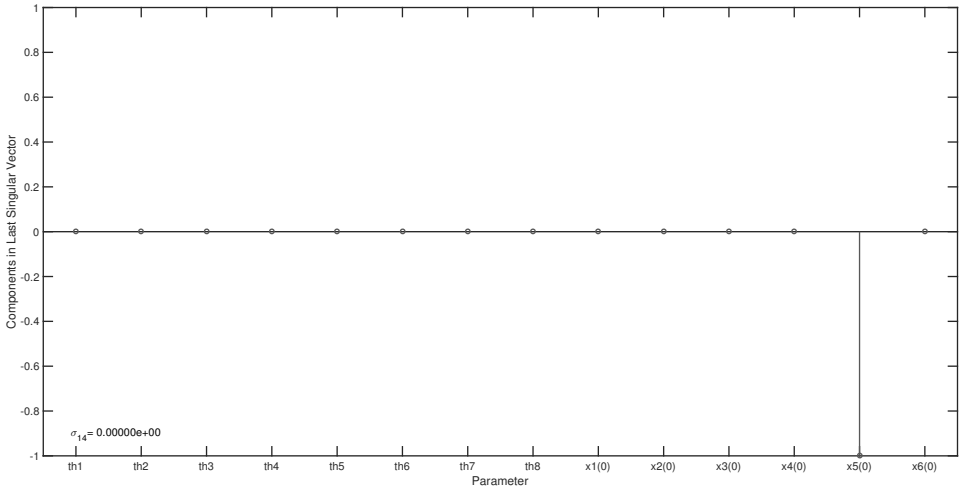


Figure 2.8: **Example 4: Ligand binding model: Entries in the last right singular vector corresponding to the smallest singular value of precisely zero, calculated for the measured output $\{x_1, x_2, x_3, x_4, x_6\}$.** The nontrivial null-space indicates that the initial condition of state x_5 is unidentifiable when this state is not measured. Accordingly, x_5 should be included into the model's minimal output set.

EXAMPLE 5: SIMPLIFIED GLYCOLYTIC REACTION MODEL

The simplified glycolytic reaction map consists of 10 chemical species: glucose, ADP, glucose 6-phosphate, ATP, glucose 1-phosphate, AMP, fructose 6-phosphate, fructose 2,

6-biphosphate, triose phosphate and pyruvate. The interaction between these chemicals are described by 9 reactions [27] (see supplementary S5 File in the appendix). This model's minimal output set for observability was defined by Liu and co-authors as $\{x_{10}\}$ [6]. Our algorithm confirmed that this minimal set also ensures the model's structural identifiability. This result was obtained after 2 minutes and 43 seconds. The set of unidentifiable parameters, $\phi_1 = \{\theta_{13}, x_{10}(0)\}$, corresponding with the omitted sensor set, $\psi_1 = \{x_{10}\}$, is indicated in Fig 2.9.

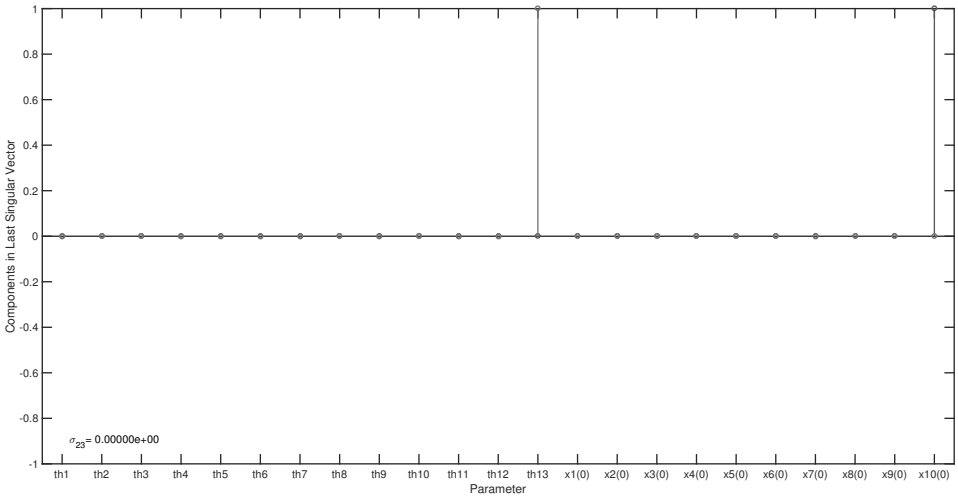


Figure 2.9: **Example 5: Simplified glycolytic reaction model: Entries in the right singular vectors corresponding to 2 vanishing singular values.** The nonzero values indicate that the initial condition $x_{10}(0)$ and parameter θ_{13} are unidentifiable when state x_{10} is not measured.

EXAMPLE 6: GOLDBETER MODEL

Consider a model describing the circadian oscillations in the Drosophila period protein (PER) [28]. It is based on both multiple phosphorylation of PER and on the negative feedback exerted by PER on the transcription of the period (per) gene. It provides a molecular basis for circadian oscillations of the limit cycle type in which the peak in per mRNA precedes the peak in total PER protein.

This model was analysed by Sedoglavic in 1995, in which he identified only 1 set of totally correlated parameters [29]. It contains 5 states and 17 system parameters and assuming that initial conditions are unknown, the total number of unknown parameters is 22. Measuring the output, $\{x_2, x_3, x_4, x_5\}$, our algorithm also found only the 1 totally correlated set, $\phi_1 = \{\theta_1, \theta_3, \theta_4, \theta_5, x_1(0)\}$, with its elements indicated by the nonzero values in Fig 2.10. The minimal output set of this model, $\{x_1\}$, was calculated in 12 seconds.

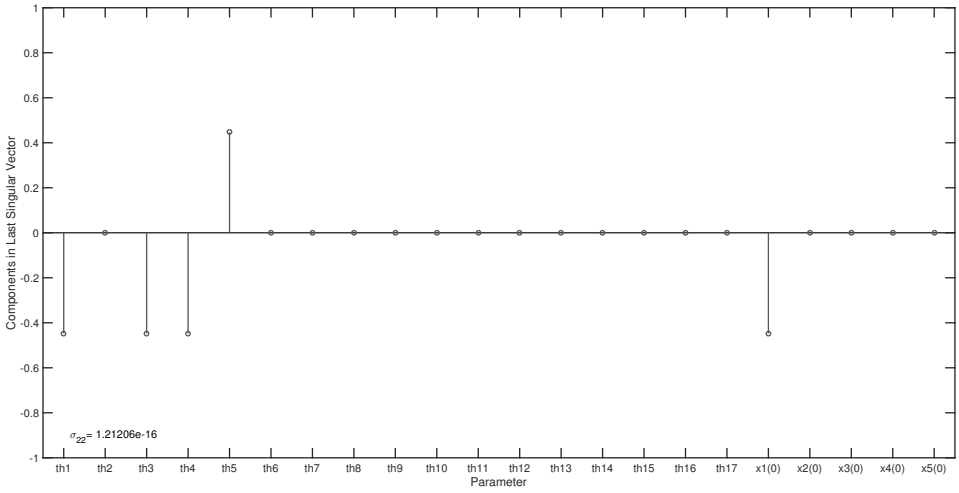


Figure 2.10: **Example 6: Goldbeter model: Entries in the last right singular vector corresponding to a single vanishing singular value calculated.** The nonzero values indicate that parameters $\theta_1, \theta_3, \theta_4, \theta_5$ and initial condition $x_1(0)$ are unidentifiable when state x_1 is not measured.

EXAMPLE 7: REPARAMETERISED JAK/STAT MODEL WITH SPECIFIC MODEL OUTPUT

In this example, we illustrate how our method can be used to identify minimal output sets from a set of more complex model outputs. These outputs do not simply consist of single model states and in this example, also include additional unknown model parameters. We consider a reparameterised JAK/STAT model, with the original unidentifiable model described by Raia and co-authors [30]. The constitutive activation of the JAK (Janus kinase)/STAT signalling pathway forms part of both the primary mediastinal B-cell lymphoma (PMBL) and the classical Hodgkin lymphoma (cHL). Raue and co-authors investigated the identifiability of this benchmark model using three different approaches [2].

The model definition also contains a specific set of initial conditions for model states, $\boldsymbol{x}(0) = \{1.3, \theta_{21}, 0, 1, 0, 2.8, 0, 165, 0, 0, 0.34, 0, 0, 0\}$. These initial conditions, in conjunction with the predetermined set of model outputs, result in the model's structural unidentifiability. Structural identifiability can be reinstated by reparameterising the model (See supplementary S7 file for the structurally identifiable version of this JAK/STAT model). The reparameterised model contains 14 states and 21 parameters, with only the initial condition of state x_2 assumed to be unknown.

Considering the reparameterised model output, $\boldsymbol{y}_{max} = [x_1 + x_3 + x_4, \theta_{16}(x_3 + x_4 + x_5 + x_{12}), \theta_{17}(x_4 + x_5), \theta_{18}x_7, \theta_{19}x_{10}, \theta_{20}x_{14}, x_{13}, x_9]$, our algorithm can now be implemented to determine the model's minimal output sets. Setting $k = 1$, already revealed 6 essential sensors. The unidentifiable parameters obtained were: $\phi_1 = \{\theta_{12}, \theta_{16}\}$, when sensor $\psi_1 = \{\theta_{16}(x_3 + x_4 + x_5 + x_{12})\}$ was not measured, $\phi_2 = \{\theta_{17}\}$, when $\psi_2 = \{\theta_{17}(x_4 + x_5)\}$ was not measured, $\phi_3 = \{\theta_{18}\}$, when $\psi_3 = \{\theta_{18}x_7\}$ was not measured, $\phi_4 = \{\theta_{19}\}$, when $\psi_4 = \{\theta_{19}x_{10}\}$ was not measured, $\phi_5 = \{\theta_{20}\}$, when $\psi_5 = \{\theta_{20}x_{14}\}$ was not measured, and

$\phi_6 = \{\theta_8, \theta_{13}\}$ when state $\psi_6 = \{x_{13}\}$ was not measured. All these sensors are essential and the resulting minimal output set, obtained after 18 seconds, is: $\{\theta_{16}(x_3 + x_4 + x_5 + x_{12}), \theta_{17}(x_4 + x_5), \theta_{18}x_7, \theta_{19}x_{10}, \theta_{20}x_{14}, x_{13}\}$.

Figs 2.11 and 2.12 reveal the identifiability signature obtained when sensor $\theta_{17}(x_4 + x_5)$ was not measured. From this, one can see that parameter θ_{17} is unidentifiable.

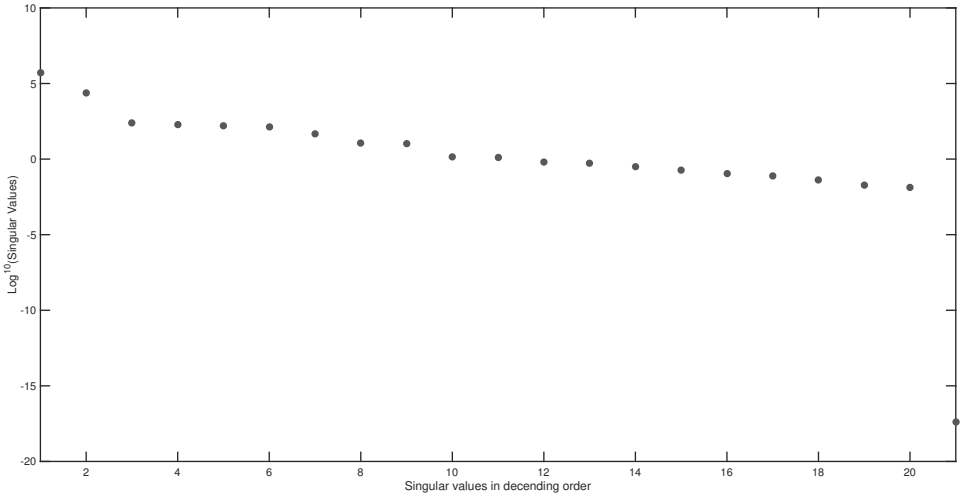


Figure 2.11: **Example 7: JAK/STAT model with specific model output: Singular values of the output sensitivity matrix, S, when omitting sensor $\theta_{17}(x_4 + x_5)$ from y_{max}.** Singular values, arranged in descending order, reveal a clear gap. This gap in conjunction with the smallest singular value of 4×10^{-18}, indicate that S is rank deficient.

EXAMPLE 8: IRRATIONAL JAK/STAT MODEL WITH SPECIFIC MODEL OUTPUT

In this final example, we show that our method can be used to analyse irrational models. Consider a irrational version of the JAK/STAT model in example 7:

$$\dot{x}_1 = \theta_1 c_1 u_1 x_1 - \theta_2 x_1^{\pi} + \theta_3 x_2, \tag{2.10}$$

$$\dot{x}_2 = \theta_2 x_1 - \theta_3 x_2, \tag{2.11}$$

$$\dot{x}_3 = \theta_1 c_1 u_1 x_1 - \theta_4 x_3 x_7, \tag{2.12}$$

$$\dot{x}_4 = \theta_4 x_3 x_7 - \theta_5 x_4, \tag{2.13}$$

$$\dot{x}_5 = \theta_5 x_4 - \theta_6 x_5, \tag{2.14}$$

$$\dot{x}_6 = -\frac{\theta_7 x_3 x_6}{1+\theta_8 x_{13}} - \frac{\theta_7 x_4 x_6^{\sqrt{2}}}{1+\theta_8 x_{13}} + c_2\theta_9 x_7, \tag{2.15}$$

$$\dot{x}_7 = \frac{\theta_7 x_3 x_6}{1+\theta_8 x_{13}} + \frac{\theta_7 x_4 x_6^{\sqrt{2}}}{1+\theta_8 x_{13}} - c_2\theta_9 x_7, \tag{2.16}$$

$$\dot{x}_8 = -\theta_{10} x_8 x_7 + c_2\theta_{11} x_9, \tag{2.17}$$

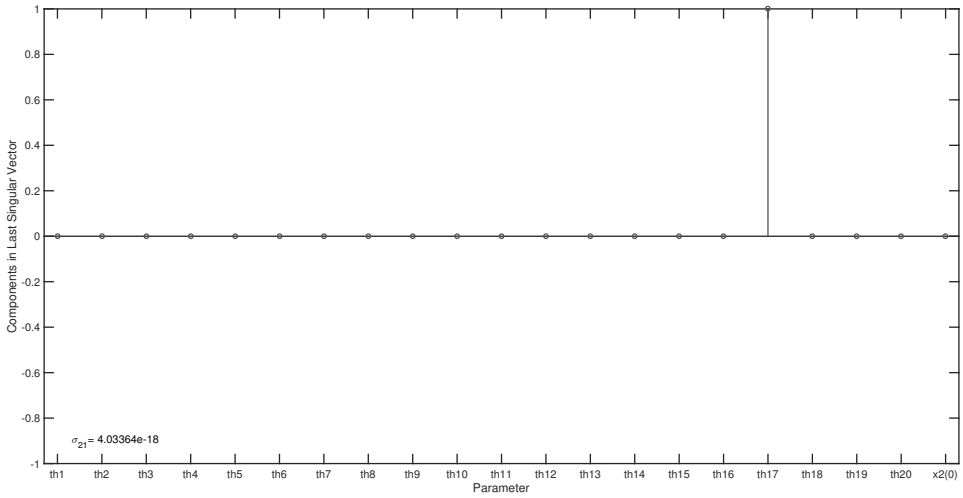


Figure 2.12: **Example 7: JAK/STAT model with specific model output: Entries in the last right singular vector corresponding to the vanishing singular value in Fig 2.11.** The nontrivial null-space indicates that model parameter θ_{17} is not identifiable when sensor $\theta_{17}(x_4 + x_5)$ is not measured.

$$\dot{x}_9 = \theta_{10} x_8 x_7 - c_2 \theta_{11} x_9, \tag{2.18}$$

$$\dot{x}_{10} = x_9, \tag{2.19}$$

$$\dot{x}_{11} = -\theta_{12} c_1 u_1 x_{11}, \tag{2.20}$$

$$\dot{x}_{12} = \theta_{12} c_1 u_1 x_{11}, \tag{2.21}$$

$$\dot{x}_{13} = \frac{\theta_{13} \sin(x_{10})}{\theta_{14} + x_{10}} - \theta_{15} x_{13}, \tag{2.22}$$

$$\dot{x}_{14} = x_9. \tag{2.23}$$

Analysing this model, we find the results identical to those obtained in example 7 and therefore conclude that the predefined outputs, $x_1 + x_3 + x_4$ and x_9, do not have to be measured to ensure this model's identifiability.

Model descriptions can be found in the supplementary material (see supplementary files S1 File to S7 file). The symbolic verification of the individual unidentifiable sets in $\boldsymbol{\phi}$ can be found in the supplementary S8 File. The MATLAB code of our algorithm can be found at: https://sourceforge.net/u/djoubert-wur/profile.

2.4. CONCLUSIONS

In this paper we introduced an algorithm that can find minimal output sets for a wide range of models in a short time. It is not limited by any specific model structure. Proposing multiple plausible minimal output sets to experimental researchers enables them to select model outputs based on factors such as measurement cost and complexity. Offering measurement flexibility whilst ensuring structural identifiability is a useful tool to scientists and our algorithm could propose these minimal sets within a couple of min-

utes.

In the future we intent to increase the numerical accuracy of our method, potentially making use of the increased integration accuracy obtained by using complex derivatives to compute matrices $\partial \boldsymbol{f}/\partial \boldsymbol{x}$ and $\partial \boldsymbol{f}/\partial \boldsymbol{\theta}$. This step could increase the tolerance of the elements of the output sensitivity matrix to 10^{-20} [31]. In addition, we are investigating the added advantages of concatenating the sensitivity matrix for different sets of the model parameter values. Preliminary results indicate that this can have a dramatic effect on the accuracy in our computations [23].

REFERENCES

[1] D. Joubert, J. Stigter, and J. Molenaar, *Determining minimal output sets that ensure structural identifiability,* PLoS One **13**, e0207334 (2018).

[2] A. Raue, J. Karlsson, M. P. Saccomani, M. Jirstrand, and J. Timmer, *Comparison of approaches for parameter identifiability analysis of biological systems,* Bioinformatics **30**, 1440–1448 (2014).

[3] D. J. Cole, B. J. T. Morgan, and D. M. Titterington, *Determining the parametric structure of models,* Mathematical Biosciences **228**, 16 (2010).

[4] M. Anguelova, J. Karlsson, and M. Jirstrand, *Minimal output sets for identifiability,* Mathematical Biosciences **239**, 139 (2012).

[5] E. August, *Parameter identifiability and optimal experimental design,* (IEEE Computational Society, 2009) pp. 29–31.

[6] Y. Liu, J. Slotine, and B. A.L., *Observability of complex systems,* Proc Natl Acad Sci **10**, 2460 (2013).

[7] C. Letellier, I. Sendiña Nadal, E. Bianco-Martinez, and M. Baptista, *A symbolic network-based nonlinear theory for dynamical systems observability,* Sci Rep **8**, 1 (2018).

[8] E. T. Tunali and T. J. Tarn, *New results for identifiability of nonlinear systems,* IEEE Transactions on Automatic Control **32**, 146 (1987).

[9] L. C. R. R. Cormen, T.H. and C. Stein, *Introduction to Algorithms* (Cambridge: The MIT Press, 2009).

[10] P. Docherty, J. Chase, T. Lotz., and T. Desaive, *A graphical method for practical and informative identifiability analyses of physiological models: A case study of insulin kinetics and sensitivity,* Biomed Eng Online **10**, 39 (2011).

[11] J. D. Stigter and J. Molenaar, *A fast algorithm to assess local structural identifiability,* Automatica **58**, 118 (2015).

[12] B. J. Chis, O.T. and E. Balsa-Canto, *Structural identifiability of systems biology models: A critical comparison of methods,* PLoS ONE **6**, e27755 (2011).

[13] J. Reid, *Structural identifiability in linear time invariant systems,* IEEE Transactions on Automatic Control **22**, 242 (1977).

[14] A. Cintrón-Arias, H. Banks, A. Capaldi, and L. A.L., *A sensitivity matrix based methodology for inverse problem formulation,* Journal of Inverse and Ill-posed Problems **17**, 545 (2009).

[15] C. Cobelli and J. DiStefano-III, *Parameter and structural identifiability concepts and ambiguities: a critical review and analysis,* Am J Physiol **239**, R7 (1980).

[16] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th ed. (The Johns Hopkins University Press, 2013).

[17] F. Gantmacher, *The Theory of Matrices* (New York: Chelsea publishing company, 1960).

[18] Y. Bard, *Nonlinear Parameter Estimation* (Academic Press Inc, 1974).

[19] H. Miao, X. Xia, A. Perelson, and H. Wu, *On identifiability of nonlinear ode models and applications in viral dynamics,* SIAM Rev Soc Ind Appl Math **53**, 3 (2011).

[20] R. Bapat, *Linear Algebra and Linear Models* (New York: Springer-Verlag, 2012).

[21] Y. Chu and J. Hahn, *Parameter set selection for estimation of nonlinear dynamic systems,* AIChE J **53**, 2858–2870 (2007).

[22] G. Quintana-Ortí and E. S. Quintana-Ortí, *Parallel codes for computing the numerical rank,* Linear Algebra and its Applications **275-276**, 451 (1998).

[23] J. D. Stigter, D. Joubert, and J. Molenaar, *Observability of complex systems: Finding the gap,* Scientific Reports **7**, 1 (2017).

[24] T. Lipniacki, P. Paszek, A. Brasier, B. Luxon, and M. Kimmel, *Mathematical model of nf-kb regulatory module,* Theor Biol **228**, 195 (2004).

[25] S. Yamada, S. Shiono, A. Joo, and A. Yoshimura, *Control mechanism of jak/stat signal transduction pathway,* FEBS Lett **534**, 190 (2003).

[26] A. Raue, V. Becker, U. Klingmüller, and J. Timmer, *Identifiability and observability analysis for experimental design in nonlinear dynamical models,* Chaos **20**, 045105 (2010).

[27] R. Heinrich and S. Schuster, *The regulation of Cellular Systems* (Berlin: Springer, 1996).

[28] A. Goldbeter, *A model for circadian oscillations in the drosophila period protein (per),* Proc Biol Sci **261**, 319 (1995).

[29] A. Sedoglavic, *A probabilistic algorithm to test local algebraic observability in polynomial time,* Journal of Symbolic Computation **33**, 735 (2002).

[30] V. Raia, M. Schilling, M. Böhm, B. Hahn, A. Kowarsch, A. Raue, C. Sticht, S. Bohl, M. Saile, P. Möller, N. Gretz, J. Timmer, F. Theis, W.-D. Lehmann, P. Lichter, and U. Klingmüller, *Dynamic mathematical modeling of IL13-induced signaling in hodgkin and primary mediastinal B-cell lymphoma allows prediction of therapeutic targets,* Cancer Research **71**, 693 (2011).

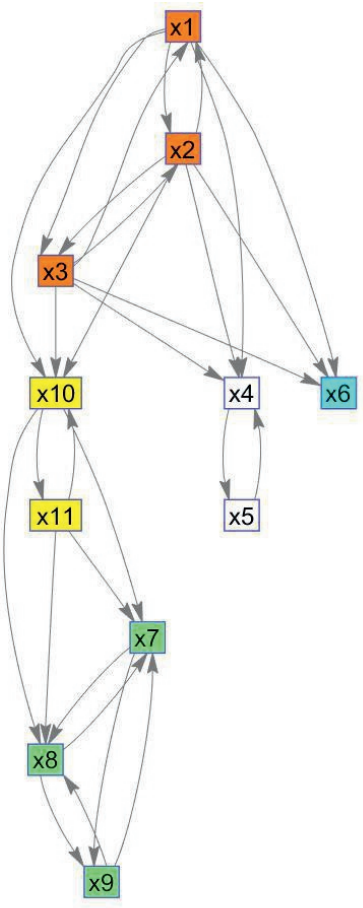
[31] J. Martins, P. Sturdza, and J. Alonso, *The connection between the complex-step derivative approximation and algorithmic differentiation,* in *Proceedings of the 39th Aerospace Sciences Meeting* (2001).

APPENDIX

2

S1 File. A chemical reaction system description.
A description of model kinetics and all model states and parameters.
Model kinetics:

```
dx1/dt = -θ1*x1*x2*x3;
dx2/dt = -θ1*x1*x2*x3;
dx3/dt = -θ1*x1*x2*x3;
dx4/dt = θ1*x1*x2*x3 - θ2*x4 + θ3*x5;
dx5/dt = θ2*x4 - θ3*x5;
dx6/dt = θ1*x1*x2*x3;
dx7/dt = θ4*x8*x9 - θ5*x7 + θ6*x10*x11;
dx8/dt = -θ4*x8*x9 + θ5*x7 + θ6*x10*x11;
dx9/dt = -θ4*x8*x9 + θ5*x7;
dx10/dt = θ1*x1*x2*x3 - θ6*x10*x11;
dx11/dt = -θ6*x10*x11
```



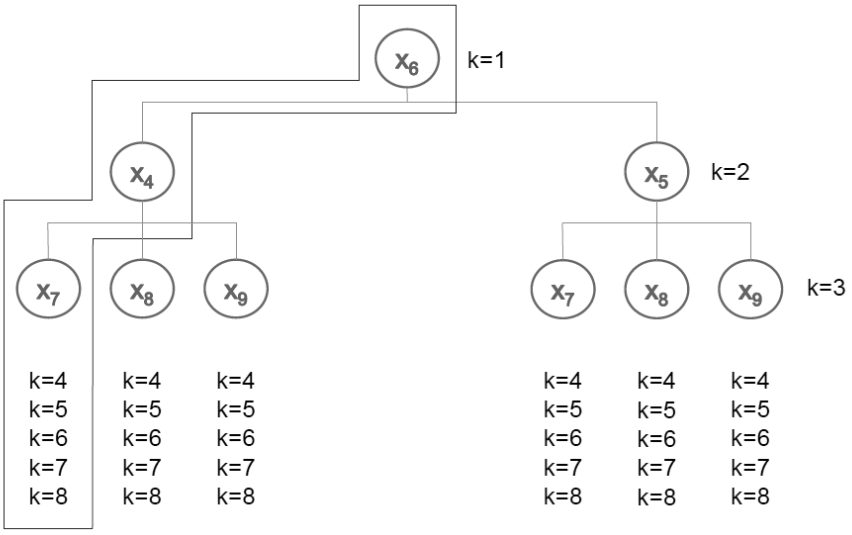
Initial conditions as additional model parameters:

θ_7	$x_1(0)$
θ_8	$x_2(0)$
θ_9	$x_3(0)$
θ_{10}	$x_4(0)$
θ_{11}	$x_5(0)$
θ_{12}	$x_6(0)$
θ_{13}	$x_7(0)$
θ_{14}	$x_8(0)$
θ_{15}	$x_9(0)$
θ_{16}	$x_{10}(0)$
θ_{17}	$x_{11}(0)$

Model output containing all measurable outputs:

$\mathbf{y}_{max} = [x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}]$

MINIMAL OUTPUT SET ALGORITHM – BRANCHING OF ANALYSIS ILLUSTRATED GRAPHICALLY



For k=4,...,8 - 6 different branches, with an output always containing the essential sensors indicated in the graph, has to be analysed.

S2 File. NF-κB model description.

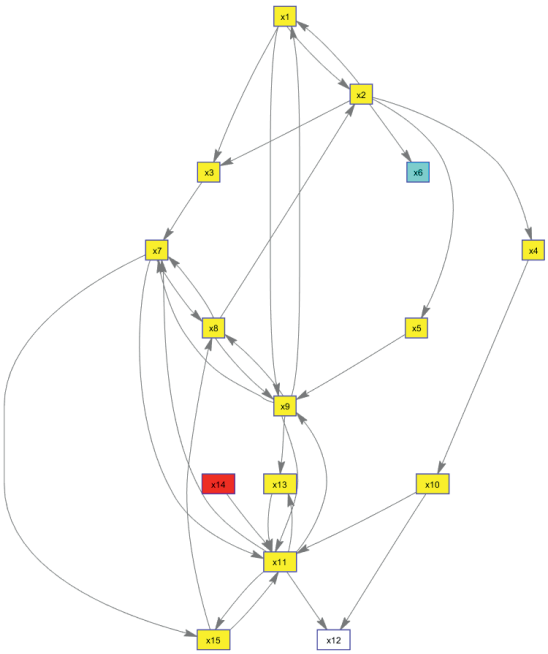
A description of model kinetics and all model states and parameters.

Model kinetics:

```
dx1/dt= -01*x1*x2+((1/333)*(-014*x1+015*x9));
dx2/dt= -01*x1*x2+(1/333)*(013*x8);
dx3/dt= 01*x1*x2-(1/333)*(011*x3);
dx4/dt= 03+02*x2-04*x4;
dx5/dt= 06+05*x2-07*x5;
dx6/dt= 09+08*x2-010*x6;
dx7/dt= (10/16667)*10*011*x3-021*x7+01*x8*x9-028*x7*x11;
dx8/dt= 021*x7-(10/16667)*013*x8-01*x8*x9+026*x15;
dx9/dt= 018*x5-023*x9-01*x8*x9+(10/16667)*(014*x1-015*x9)-025*x9*x11;
dx10/dt= 027*x4-024*x10;
dx11/dt= -012*x11-016*x11-028*x7*x11-025*x9*x11-019*x10*x11+017*x13+...
022*x14+026*x15;
dx12/dt= 016*x11+019*x10*x11-012*x12;
dx13/dt= 025*x9*x11-017*x13;
dx14/dt= 020-012*x14-022*x14;
dx15/dt= 028*x7*x11-026*x15
```

Initial conditions as additional model parameters:

θ_{29}	$x_1(0)$
θ_{30}	$x_2(0)$
θ_{31}	$x_3(0)$
θ_{32}	$x_4(0)$
θ_{33}	$x_5(0)$
θ_{34}	$x_6(0)$
θ_{35}	$x_7(0)$
θ_{36}	$x_8(0)$
θ_{37}	$x_9(0)$
θ_{38}	$x_{10}(0)$
θ_{39}	$x_{11}(0)$
θ_{40}	$x_{12}(0)$
θ_{41}	$x_{13}(0)$
θ_{42}	$x_{14}(0)$
θ_{43}	$x_{15}(0)$



Model output containing all measurable outputs:

$\mathbf{y}_m = [x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}]$

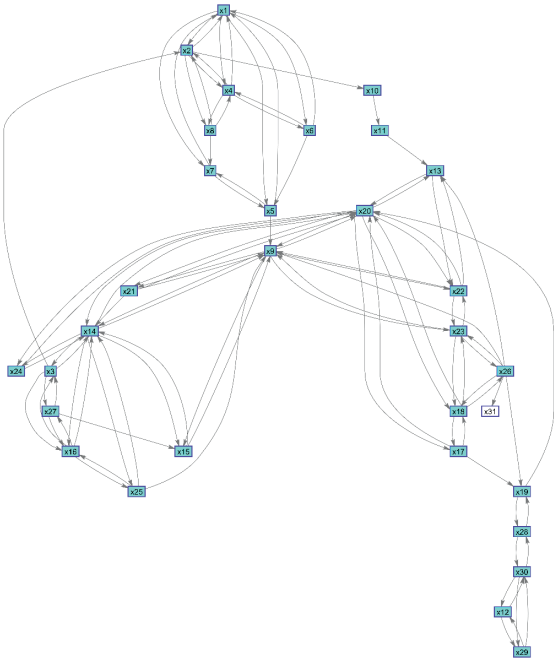
S3 File. JAK-STAT model description.
A description of model kinetics and all model states and parameters.
Model kinetics:

```
dx1/dt = -2*(x1*x1*01 - x2*02) - x1*x4*04 + x6*05 - x1*x5*07 + x7*08;
dx2/dt = x1*x1*01 - x2*02 + x3*03 - x2*x4*09 + x8*010;
dx3/dt = -x3*03 + x14*x14*023 - x3*024 - x3*x16*030 + x27*031;
dx4/dt = -x1*x4*04 + x6*05 + x6*06 - x2*x4*09 + x8*010 + x8*011;
dx5/dt = x6*06 - x1*x5*07 + x7*08 - x5*012;
dx6/dt = x1*x4*04 - x6*05 - x6*06;
dx7/dt = x1*x5*07 - x7*08 + x8*011;
dx8/dt = x2*x4*09 - x8*010 - x8*011;
dx9/dt = x5*012 - x9*x14*016 + x15*017 - x9*x22*019 + x23*020 - x9*x20*021 + x21*022 +
x25*029 + x26*047;
dx10/dt = -x10*013 + (x2*014)/(x2 + 015);
dx11/dt = x10*013 - x11*050;
dx12/dt = -x12*x29*041 + x30*042;
dx13/dt = -x13*x20*025 + x22*026 + x26*047 - x13*048 + x11*051;
```

```
dx14/dt = -x9*x14*016 + x15*017- 2*(x14*x14*023 - x3*024) + x21*027 - x14*x20*033 +
x24*034 - x14*x16*035 + x25*036;
dx15/dt = x9*x14*016 - x15*017 + x27*028;
dx16/dt = x27*028 + x25*029 - x3*x16*030 + x27*031 - x14*x16*035 + x25*036;
dx17/dt = -x17*018 + x18*x20*039 - x17*040;
dx18/dt = x17*018 - x18*x20*039 + x17*040 - x18*x23*045 + x26*046* + x26*047;
dx19/dt = x17*018 - x19*032 + x28*x28*037 - x19*038 + x26*047;
dx20/dt = -x9*x20*021 + x21*022 - x13*x20*025 + x22*026 + x21*027 + x19*032 - x14*x20*033
+ x24*034 - x18*x20*039 + x17*040;
dx21/dt = x9*x20*021 - x21*022 - x21*027;
dx22/dt = -x9*x22*019 + x23*020 + x13*x20*025 - x22*026;
dx23/dt = x9*x22*019 - x23*020 - x18*x23*045 + x26*046;
dx24/dt = x14*x20*033 - x24*034;
dx25/dt = -x25*029 + x14*x16*035 - x25*036;
dx26/dt = x18*x23*045 - x26*046 - x26*047 - x26*049;
dx27/dt = -x27*028 + x3*x16*030 - x27*031;
dx28/dt = -2*(x28*x28*037 - x19*038) + x30*043 - x28*044;
dx29/dt = -x12*x29*041 + x30*042;
dx30/dt = x12*x29*041 - x30*042 - x30*043 + x28*044;
dx31/dt = x26*049
```

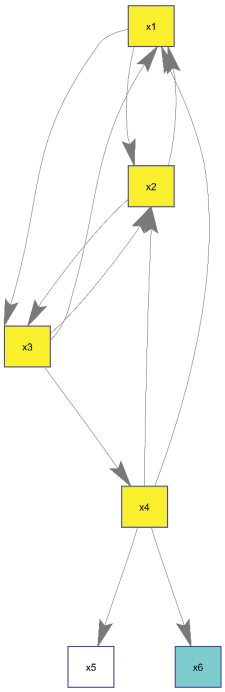
Additional model parameters: $\theta_{52}, \ldots, \theta_{82} = x_1(0), \ldots, x_{31}(0)$

Model output: $\mathbf{y}_m = [x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{17}, x_{18}, x_{19}, x_{20}, x_{21}, x_{22}, x_{23}, x_{24}, x_{25}, x_{26}, x_{27}, x_{28}, x_{29}, x_{30}, x_{31}]$



S4 File. Ligand binding model description.
A description of model kinetics and all model states and parameters.
Model kinetics:

```
dx1/dt = -θ1*x1*x2 + θ1*θ2*x3 + θ5*x4;
dx2/dt = -θ1*x1*x2 + θ1*θ2*x3 + θ3*θ8 - θ3*x2 + θ5*x4;
dx3/dt = θ1*x1*x2 - θ1*θ2*x3 - θ4*x3;
dx4/dt = θ4*x3 - θ5*x4 - θ6*x4 - θ7*x4;
dx5/dt = θ6*x4;
dx6/dt = θ7*x4
```



Model parameters:

θ_1	k_{on}	θ_9	$x_1(0)$	Epo
θ_2	k_D	θ_{10}	$x_2(0)$	EpoR
θ_3	k_t	θ_{11}	$x_3(0)$	Epo_EpoR
θ_4	k_e	θ_{12}	$x_4(0)$	Epo_EpoR_i
θ_5	k_{ex}	θ_{13}	$x_5(0)$	Epo_i
θ_6	k_{di}	θ_{14}	$x_6(0)$	Epo_e
θ_7	k_{de}			
θ_8	B_{max}			

Model output:

$\mathbf{y}_{max} = [x_1, x_2, x_3, x_4, x_5, x_6]$

S5 File. Simplified glycolytic model description.
A description of model kinetics and all model states and parameters.
Model kinetics:

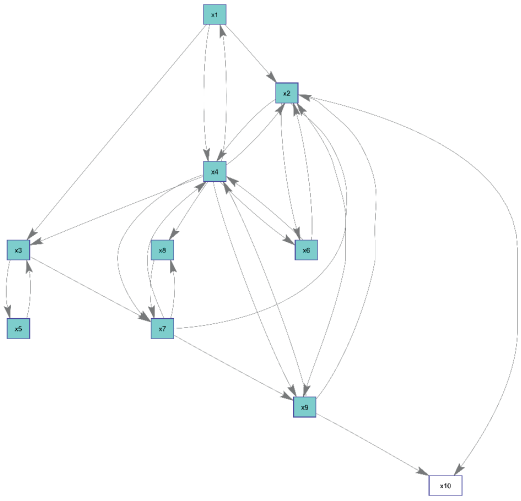
```
dx1/dt = -θ1*x1*x4 + θ12;
dx2/dt = θ1*x1*x4 + θ4*x4*x6 - θ5*x2*x2 + θ7*x4*x7 + θ9*x4*x7 + θ10*x4 - θ11*x2*x2*x9;
dx3/dt = θ1*x1*x4 - θ2*x3 + θ3*x5 - θ6*x3;
dx4/dt = -θ1*x1*x4 - θ4*x4*x6 + θ5*x2*x2 - θ7*x4*x7 - θ9*x4*x7 - θ10*x4 + θ11*x2*x2*x9;
dx5/dt = θ2*x3 - θ3*x5;
dx6/dt = -θ4*x4*x6 + θ5*x2*x2;
dx7/dt = θ6*x3 - θ7*x4*x7 + θ8*x8 - θ9*x4*x7;
dx8/dt = θ7*x4*x7 - θ8*x8;
dx9/dt = θ9*x4*x7 - θ11*x2*x2*x9;
dx10/dt = θ11*x2*x2*x9 - θ13*x10
```

Additional model parameters:

θ_{12}	C_1	θ_{14}	$x_1(0)$	glucose
θ_{13}	C_2	θ_{15}	$x_2(0)$	ADP
		θ_{16}	$x_3(0)$	G6P
		θ_{17}	$x_4(0)$	ATP
		θ_{18}	$x_5(0)$	G1P
		θ_{19}	$x_6(0)$	AMP
		θ_{20}	$x_7(0)$	F6P
		θ_{21}	$x_8(0)$	F2-6BP
		θ_{22}	$x_9(0)$	TP
		θ_{23}	$x_{10}(0)$	Pyr

Model output:

$\mathbf{y}_m = [x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}]$

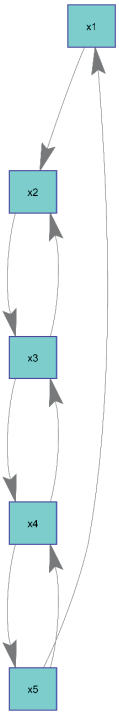


S6 File. Goldbeter model with specific model output description.
A description of model kinetics and all model states and parameters.
Model kinetics:

```
dx1/dt = (th1*th2^4)/(th2^4+x5^4) - (th3*x1)/(th4+x1);
dx2/dt = th5*x1-(th6*x2)/(th7+x2)+(th8*x3)/(th9+x3);
dx3/dt = (th6*x2)/(th7+x2)+(th10*x4)/(th11+x4)-x3*((th8)/(th9+x3)+(th12)/(th13+x3));
dx4/dt = (th12*x3)/(th13+x3)-x4*((th10)/(th11+x4)+th14+(th15)/(th16+x4))+th17*x5;
dx5/dt = th14*x4-th17*x5
```

Initial conditions as additional model parameters:

θ_{18}	$x_1(0)$	M
θ_{19}	$x_2(0)$	P_0
θ_{20}	$x_3(0)$	P_1
θ_{21}	$x_4(0)$	P_2
θ_{22}	$x_5(0)$	P_N



Defined model parameters:

θ_1	v_s
θ_2	K_I
θ_3	v_m
θ_4	K_m
θ_5	k_s
θ_6	V_1
θ_7	K_1
θ_8	V_2
θ_9	K_2
θ_{10}	V_4
θ_{11}	K_4
θ_{12}	V_3
θ_{13}	K_3

θ_{14}	k_1
θ_{15}	v_d
θ_{16}	K_d
θ_{17}	k_2

Model output containing all measurable outputs:

$\mathbf{y}_m = [x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}]$

S7 File. Re-parametrised JAK-STAT model description.
A description of model kinetics and all model states and parameters.

Model kinetics:

```
dx1/dt = θ1*c1*u1*x1-θ2*x1+θ3*x2;
dx2/dt = θ2*x1-θ3*x2;
dx3/dt = θ1*c1*u1*x1-θ4*x3*x7;
dx4/dt = θ4*x3*x7-θ5*x4;
dx5/dt = θ5*x4-θ6*x5;
dx6/dt = -θ7*x3*x6/(1+θ8*x13)-θ7*x4*x6/(1+θ8*x13)+c2*θ9*x7;
dx7/dt = θ7*x3*x6/(1+θ8*x13)+θ7*x4*x6/(1+θ8*x13)-c2*θ9*x7;
dx8/dt = -θ10*x8*x7+c2*θ11*x9;
dx9/dt = θ10*x8*x7-c2*θ11*x9;
dx10/dt = x9;
dx11/dt = -θ12*c1*u1*x11;
dx12/dt = θ12*c1*u1*x11;
dx13/dt = θ13*x10/(θ14+x10)-θ15*x13;
dx14/dt = x9;
```

Constants:

c1=2.265;
c2=1;
u1=4;

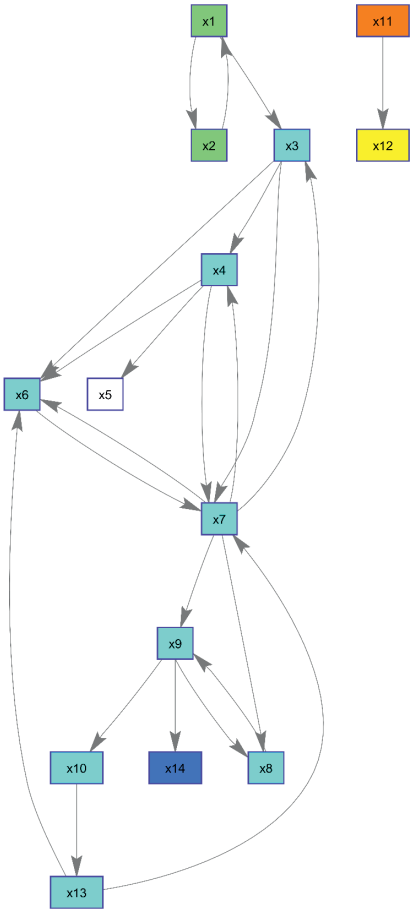
Model parameters and initial conditions:

	Initial conditions
$x_1(0)$	1.3
$x_2(0)$	θ_{21}
$x_3(0)$	0
$x_4(0)$	1

$x_5(0)$	0
$x_6(0)$	2.8
$x_7(0)$	0
$x_8(0)$	165
$x_9(0)$	0
$x_{10}(0)$	0
$x_{11}(0)$	0.34
$x_{12}(0)$	0
$x_{13}(0)$	0
$x_{14}(0)$	0

Model output:

$\mathbf{y_m}$ = [x1+x3+x4;
θ16*(x3+x4+x5+x12);
θ17*(x4+x5);
θ18*x7;
θ19*x10;
θ20*x14;
x13;
x9]



Determining the minimal output sets that ensure the structural identifiability of a model

S8 File. Symbolically verified sets of correlated parameters.

This document contains symbolic verification of some of the totally correlated parameter sets identified in this paper. Refer to Stigter and Molenaar for details regarding these computations [1].

Example 1: A chemical reaction system

- When state x_6 is not measured
  Non-trivial null-space computed: $\mathcal{N}\left(\boldsymbol{J}(\tilde{\boldsymbol{x}}_0^{unid})\right) = \{1\}$. Here $\tilde{\boldsymbol{x}}_0^{unid} = \{x_6(0)\}$
- When states x_4 and x_5 are not measured
  Non-trivial null-space computed:
  $\mathcal{N}\left(\boldsymbol{J}(\tilde{\boldsymbol{x}}_0^{unid})\right) = \{1,0,0,0\}, \{0,1,0,0\}, \{0,0,1,0\}, \{0,0,0,1\}$. Here
  $\tilde{\boldsymbol{x}}_0^{unid} = \{\theta_2, \theta_3, x_4(0), x_5(0)\}$.
- When states x_7, x_8 and x_9 are not measured
  Non-trivial null-space computed: $\mathcal{N}\left(\boldsymbol{J}(\tilde{\boldsymbol{x}}_0^{unid})\right) =$
  $\{1,0,0,0,0\}, \{0,1,0,0,0\}, \{0,0,1,0,0\}, \{0,0,0,1,0\}, \{0,0,0,0,1\}$. Here
  $\tilde{\boldsymbol{x}}_0^{unid} = \{\theta_4, \theta_5, x_7(0), x_8(0), x_9(0)\}$.

Example 2: NF-κB model

- When state x_4 is not measured
  Non-trivial null-space computed:
  $\mathcal{N}\left(\boldsymbol{J}(\tilde{\boldsymbol{x}}_0^{unid})\right) = \{\theta_2/x_4(0), -\theta_3/x_4(0), -\theta_{27}/x_4(0), 1\}$, were
  $\tilde{\boldsymbol{x}}_0^{unid} = \{\theta_2, \theta_3, \theta_{27}, x_4(0)\}$.

- When state x_5 is not measured

  Non-trivial null-space computed:
  $\mathcal{N}\left(\boldsymbol{J}(\tilde{\boldsymbol{x}}_0^{unid})\right) = \{\theta_5/x_5(0), \theta_6/x_5(0), -\theta_{18}/x_5(0), 1\}$, were
  $\tilde{\boldsymbol{x}}_0^{unid} = \{\theta_5, \theta_6, \theta_{18}, x_5(0)\}$.

- When state x_6 is not measured

  Non-trivial null-space computed:
  $\mathcal{N}\left(\boldsymbol{J}(\tilde{\boldsymbol{x}}_0^{unid})\right) = \{0, 0, 0, 1\}, \{0, 0, 1, 0\}, \{0, 1, 0, 0\}, \{1, 0, 0, 0\}$, were
  $\tilde{\boldsymbol{x}}_0^{unid} = \{\theta_8, \theta_9, \theta_{10}, x_6(0)\}$.

- When state x_{10} is not measured

  Non-trivial null-space computed:
  $\mathcal{N}\left(\boldsymbol{J}(\tilde{\boldsymbol{x}}_0^{unid})\right) = \{-\theta_{19}/x_{10}(0), \theta_{27}/x_{10}(0)\}$, were $\tilde{\boldsymbol{x}}_0^{unid} = \{\theta_{19}, \theta_{27}, x_{10}(0)\}$.

- When state x_{12} is not measured

  Non-trivial null-space computed: $\mathcal{N}\left(\boldsymbol{J}(\tilde{\boldsymbol{x}}_0^{unid})\right) = \{1\}$, were $\tilde{\boldsymbol{x}}_0^{unid} = \{x_{12}(0)\}$.

Example 3: JAK/STAT model

- When state x_{31} is not measured

  Non-trivial null-space computed: $\mathcal{N}\left(\boldsymbol{J}(\tilde{\boldsymbol{x}}_0^{unid})\right) = \{1\}$. Here $\tilde{\boldsymbol{x}}_0^{unid} = \{x_{31}(0)\}$

- When states x_{10} and x_{11} are not measured

  Non-trivial null-space computed:
  $\mathcal{N}\left(\boldsymbol{J}(\tilde{\boldsymbol{x}}_0^{unid})\right) = \{\theta_{14}/x_{11}(0), -\theta_{51}/x_{11}(0), \theta_{10}/x_{11}(0), 1\}$. Here
  $\tilde{\boldsymbol{x}}_0^{unid} = \{\theta_{14}, \theta_{51}, x_{10}(0), x_{11}(0)\}$.

Example 4: Ligand binding model

- When state x_5 is not measured

  Non-trivial null-space computed: $\mathcal{N}\left(\boldsymbol{J}(\tilde{\boldsymbol{x}}_0^{unid})\right) = \{1\}$. Here $\tilde{\boldsymbol{x}}_0^{unid} = \{x_5(0)\}$

- When state x_6 is not measured

  Non-trivial null-space computed: $\mathcal{N}\left(\boldsymbol{J}(\tilde{\boldsymbol{x}}_0^{unid})\right) = \{1\}$. Here $\tilde{\boldsymbol{x}}_0^{unid} = \{x_6(0)\}$.

Example 5: Simplified glycolytic reaction model

- When state x_{10} is not measured

Non-trivial null-space computed: $\mathcal{N}\left(\boldsymbol{J}(\tilde{\boldsymbol{x}}_0^{unid})\right) = \{1,0\}\{0,1\}$. Here $\tilde{\boldsymbol{x}}_0^{unid} = \{\theta_{13}, x_{10}(0)\}$.

Example 6: Goldbeter model

- When state x_4 is not measured

  Non-trivial null-space computed:
  $\mathcal{N}\left(\boldsymbol{J}(\tilde{\boldsymbol{x}}_0^{unid})\right) = \{\theta_1/x_1(0), \theta_3/x_1(0), \theta_4/x_1(0), -\theta_5/x_1(0), 1\}$, were $\tilde{\boldsymbol{x}}_0^{unid} = \{\theta_1, \theta_3, \theta_4, \theta_5, x_1(0)\}$.

Example 7: JAK-STAT model with specific model output

- When output $\theta_{16}(x_3 + x_4 + x_5 + x_{12})$ is not measured

  Non-trivial null-space computed: $\mathcal{N}\left(\boldsymbol{J}(\tilde{\boldsymbol{x}}_0^{unid})\right) = \{\theta_{12}, 0\}\{0, \theta_{16}\}$, were $\tilde{\boldsymbol{x}}_0^{unid} = \{\theta_{12}, \theta_{16}\}$.

- When output $\theta_{17}(x_4 + x_5)$ is not measured

  Non-trivial null-space computed: $\mathcal{N}\left(\boldsymbol{J}(\tilde{\boldsymbol{x}}_0^{unid})\right) = \{1\}$, were $\tilde{\boldsymbol{x}}_0^{unid} = \{\theta_{17}\}$.

- When output $\theta_{18}x_7$ is not measured

  Non-trivial null-space computed: $\mathcal{N}\left(\boldsymbol{J}(\tilde{\boldsymbol{x}}_0^{unid})\right) = \{1\}$, were $\tilde{\boldsymbol{x}}_0^{unid} = \{\theta_{18}\}$.

- When output $\theta_{19}x_{10}$ is not measured

  Non-trivial null-space computed: $\mathcal{N}\left(\boldsymbol{J}(\tilde{\boldsymbol{x}}_0^{unid})\right) = \{1\}$, were $\tilde{\boldsymbol{x}}_0^{unid} = \{\theta_{19}\}$.

- When output $\theta_{20}x_{14}$ is not measured

  Non-trivial null-space computed: $\mathcal{N}\left(\boldsymbol{J}(\tilde{\boldsymbol{x}}_0^{unid})\right) = \{1\}$, were $\tilde{\boldsymbol{x}}_0^{unid} = \{\theta_{20}\}$.

- When output x_{13} is not measured

  Non-trivial null8-space computed: $\mathcal{N}\left(\boldsymbol{J}(\tilde{\boldsymbol{x}}_0^{unid})\right) = \{1,0\}\{0,1\}$, were $\tilde{\boldsymbol{x}}_0^{unid} = \{\theta_{12}, \theta_{13}\}$.

References

1. Stigter JD, Molenaar J. A fast algorithm to assess local structural identifiability. Automatica. 2015; 58:118-124 doi: 10.1016/j.automatica.2015.05.004

Determining the minimal output sets that ensure the structural identifiability of a model

S9 File. Bernoulli trials: how to ensure that a set of unidentifiable parameters is identified with 99.5% certainty.

In our algorithm we mention that an exhaustive search (ER) for the different sets of essential sensors of large models may be computationally demanding. In such cases, we suggest using a process of randomly omitting sensors from an output (RS) to detect these sets. The advantage of this strategy is that it significantly reduces the number of identifiability iterations required to detect a set of essential sensors, while the probability of finding this set is almost equal to one. We use the JAK/STAT model as an example.

Assume our set of available sensors $\boldsymbol{y}$, contains N elements. We can think of the K sensors that cause a lack of identifiability as K red marbles in an urn with a total of N marbles (K red and $(N-K)$ blue ones). We now draw, without replacement, n marbles from the urn (corresponding to n missing sensors from the available sensor set $\boldsymbol{y}$, $n \geq K$). The probability that we select k out of the K sensors that cause a lack of identifiability clearly follows a hyper geometric distribution, i.e.

$$P(X=k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}, \tag{S1}$$

Since lack of identifiability is only detected in case *all* K sensors are missing from the sensor set, and not just a sub-set of them, the probability of successfully detecting the complete set of K sensors is

$$P(X=K) = \frac{\binom{K}{K}\binom{N-K}{n-K}}{\binom{N}{n}} = \frac{\binom{N-K}{n-K}}{\binom{N}{n}}, \tag{S2}$$

Having established a probability of successfully detecting the sensor set ψ that causes a

lack of identifiability, we now repeat the experiment of leaving out n sensors (from the set $\boldsymbol{y}$) R times. The probability $\bar{P}_{det}$ of *not* detecting the particular sensor set ψ is given by

$$\bar{P}_{det} = (1 - P(X = K))^R \tag{S3}$$

When performing repeated random experiments, there are essentially two variables that can be manipulated by the user, namely (i) the number of missing sensors in each random trial (n), and (ii) the number of repetitions of the Bernoulli trial (R).

As an example, consider the JAK/STAT model, which has 31 states that are assumed to be measured directly and are all available in $\boldsymbol{y}_{max}$. Assume now that we have to search for an essential sensor set ψ that contains exactly $K = 10$ sensors. The probability of detecting this set when omitting 10 sensors from $\boldsymbol{y}_{max}$ (case $n = 10$) is only 2.25×10^{-8} ($P(X = 10) = \frac{\binom{10}{10}\binom{31-10}{10-10}}{\binom{31}{10}}$), whilst the probability of identifying the set for $n = 21$ equals 7.95×10^{-3} ($P(X = 10) = \frac{\binom{10}{10}\binom{31-10}{21-10}}{\binom{31}{21}}$). The probability of detecting this particular set can further be increased by repeated random selection of sensors from $\boldsymbol{y}_{max}$, of course each time omitting 21 sensors from the available set.

More specifically, say we want do identify the set of 10 sensors with a large probability, e.g. $P_{det} = 1 - \bar{P}_{det} = 0.995$. An exhaustive search would require more than 44 million iterations, since $\binom{31}{31-10} = 44352165$. By choosing $n = 21$, we already found a successful detection to occur with a probability of 7.95×10^{-3} for one trial. Based on equation (S3), we now find that if we repeat the experiment of randomly omitting 21 sensors for a total of 644 times, then the sensor set that causes a lack of identifiability is detected with a probability of 99.5%. This, of course, is a tremendous difference in comparison with an exhaustive search.

References

1. Rice JA. Mathematical Statistics and Data Analysis. 3rd ed. Belmont: Duxbury Press; 2007.

3

AN EFFICIENT PROCEDURE TO REPARAMETERISE STRUCTURALLY UNIDENTIFIABLE MODELS

Dominique JOUBERT, Hans STIGTER, Jaap MOLENAAR

Structural identifiability analysis methods are not widely used in practice, yet, due to either the computational complexity or the lack of mature computer implementations.

(Miao, Xia, Perelson and Wu, 2011)

Under review.

ABSTRACT

AN efficient method to reparameterise structurally unidentifiable models is introduced. It significantly reduces computational demand by combining both numerical and symbolic identifiability calculations. This hybrid approach facilitates the reparameterisation of: 1) large unidentifiable ordinary differential equation models and, 2) models where state transformations are required. A model is first assessed numerically. This is done to establish which parameters might be unidentifiable and to better understand the nature of the correlation between these redundant parameters. The numerical results are then used in symbolic calculations, tasked with computing viable reparameterisations that will ensure a model's structural identifiability. The use of the preceding numerical results notably reduces the number of symbolic calculations required. We illustrate our procedure and the reparameterisation process in detail in 4 examples: 1) an immunological model with 2 states and 7 unknown parameters 2) a batch reactor model with 2 states and 6 unknown parameters 3) a JAK/STAT model with 14 states and 23 unknown parameters and 4) a lung cancer model with 21 states and 75 unknown parameters. In addition to reparameterising these unidentifiable models, we also present modellers with alternative options available to obtain structural identifiability.

3.1. INTRODUCTION

System biology models often utilise ordinary differential equations to describe physical phenomena. These models may contain large numbers of parameters and at times also initial conditions that need to be inferred from experimental data. However, in some cases the statistical inference fails. This may be due to insufficient or low quality data, which is referred to as practical unidentifiability, or due to the inherent structure of the model, referred to as structural unidentifiability [1]. One way to characterise structurally unidentifiability is to say that at least 1 parameter has a confidence interval that spans the interval $(-\infty,\infty)$. Any form of unidentifiability, also referred to as aliasing, calls into question the predictive capacity of a model and urges its user to interpret all results with caution.

If a model is classified as structurally unidentifiable, a modeller may wish to know "how can this model be made identifiable?" In this paper, we present a method that can answer this question also for large models, in an efficient way. The central concept underpinning this paper is that *unidentifiable parameters may be divided into different subsets of totally correlated or aliased parameters* [2]. Structural identifiability can only be obtained if *the correlation between parameters in each of these sets is destroyed.* This can be done in different ways:

1. Measure additional model outputs. There may be some practical restrictions to this approach, yet a good understanding of a model's minimal output sets may result in easily obtainable structural identifiability [3]. For example, if a model has 1 set of totally correlated parameters that includes a certain state's unknown initial condition, measuring this initial condition as additional output will destroy the correlation and result in identifiability of all parameters.

2. Deduce the values of one or more unidentifiable parameters from other sources. If the value of one of the unknown parameters in a totally correlated set is known, this correlation is destroyed. To destroy all correlations, the value of one parameter from each of the different totally correlated sets must be known. A parameter value can either be obtained from literature or determined in a separate experiment. It should however be noted that even if parameter values are obtained from literature, they may still require re-calibration with experimental data and so this should always be done with caution [4].

3. Reparametrise the model to remove redundant parameters. In this approach, one parameter is eliminated from each of the different totally correlated sets. For example, if a model has P parameters and $M(< P)$ sets of totally correlated parameters, the reparametrised, structurally identifiable model will have $P - M$ parameters.

4. Start an experiment from different initial conditions. It might be that the initial values are taken from a thin set of singular points that give rise to structural unidentifiability [5]. In such a case, it may be possible to regain structural identifiability by changing the initial experimental conditions.

In this paper, we introduce a method related to point 3. As side note and to complement the discussion on dealing with structural unidentifiability, points 1, 2 and 4 are also raised when relevant. The novelty of the work presented here is that we can reparameterise large ODE models and, because our calculations allow for the inclusion of initial conditions of model states as additional unknown parameters, the obtained results may also reveal the required state transformations and so, no further complex analyses are required. For better understanding, we show the reparameterisation process in detail.

Although this topic has been covered to some extent in the past, the reparameterisation including state transformations of large models have remained elusive. Previous work on this topic includes: a) symbolic methods based on exhaustive summary [1, 6, 7] and examples including state transformations [8, 9], b) numerical methods involving the Fisher Information Matrix and profile likelihood [10] and c) hybrid methods where a Jacobi matrix is symbolically calculated and analysed numerically [1] and where a sensitivity matrix is computed numerically and analysed using a singular value decomposition (SVD) [11]. The unidentifiable parameters are subsequently fixed at nominal values and so the redundant model structure remains unchanged [11].

It should be mentioned that the reparameterisation process is never unique, and that it is up to the modeller to decide which redundant parameters to eliminate from a model. Caution should always be taken, as not all reparameterisations are biologically relevant [12].

This paper is divided as follows: section 3.2 covers the topics of structural identifiability and the identification of sets of totally correlated parameters. Section 3.3 contains 4 examples that illustrate how information regarding totally correlated parameter sets can be implemented to reparameterise models. Concluding remarks are given in section 3.4.

3.2. METHODS

MODEL DEFINITION

We begin with the definition of a typical ordinary differential equation model, regularly used in systems biology. These models often describe mass balances of certain cellular constituents and can be very detailed, containing numerous model states and vast numbers of unknown parameters which need to be inferred from experimental data. In this paper, we analyse dynamic models which can be written in the standard state-space form:

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{f}(t, \boldsymbol{x}(t), \boldsymbol{\theta}), \tag{3.1}$$

$$\boldsymbol{x}(0) = \boldsymbol{x_0}, \tag{3.2}$$

$$\boldsymbol{y}(t) = \boldsymbol{h}(\boldsymbol{x}(t), \boldsymbol{\theta}), \tag{3.3}$$

where $\boldsymbol{f}$ denotes a dynamic model structure and $\boldsymbol{h}$ the output or observation function. State variables are contained in vector $\boldsymbol{x}(t)$ ($\dim(\boldsymbol{x}) = n$), parameters in vector $\boldsymbol{\theta}$ ($\dim(\boldsymbol{\theta}) = p$) and model outputs in vector $\boldsymbol{y}(t)$ ($\dim(\boldsymbol{y}) = m$). Initial values of the model states can also be unknown and in such cases we regard them as additional unknown parameters and the resulting parameter vector, $\boldsymbol{\theta}$, then has $\dim(\boldsymbol{\theta}) = p + n$.

STRUCTURAL IDENTIFIABILITY

The following sections describe how to find suitable model reparametrisations based on the structure of the model defined in (3.1) – (3.3). This is achieved by assessing the model's structural identifiability both numerically and symbolically. The reparameterisation process involves the following steps: *Step 1:* numerical identifiability analysis, *Step 2:* symbolic identifiability calculations, *Step 3:* substitute obtained reparameterisations into the original model, and *Step 4:* re-evaluate the identifiability of the reparameterised model.

STEP 1: NUMERICAL IDENTIFIABILITY ANALYSIS

The numerical identifiability method we apply uses sensitivity based calculations of model outputs with respect to model parameters [13, 14]. Reid introduced the concept of sensitivity based identifiability analysis for linear models [15]. In his paper, he defines a sensitivity matrix, **S**, with elements depicting the sensitivities of the model output with respect to model parameters, $\partial \boldsymbol{y}/\partial \boldsymbol{\theta}$. For nonlinear models, the sensitivities of model outputs to individual model parameters can be calculated from the following equations:

$$\frac{d}{dt}\left(\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\theta}}\right) = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\theta}} + \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\theta}}, \tag{3.4}$$

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{\theta}} = \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}}\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\theta}} + \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{\theta}}. \tag{3.5}$$

One obtains $\partial \boldsymbol{y}/\partial \boldsymbol{\theta}$ as a function of time by integrating equation 3.4 and substituting the solution into 3.5. By calculating these sensitivities at discrete time points on an interval $[t_0, \ldots, t_N]$, one can construct a sensitivity matrix, $\boldsymbol{S}$. If any of the initial values of model states are unknown, their identifiability can easily be assessed by regarding them as additional parameters. In such cases, $\boldsymbol{S}$ may have up to $p+n$ columns, each related to a specific parameter, $\theta_i, i = 1, \ldots, p+n$.

$$\boldsymbol{S} = \begin{pmatrix} \frac{\partial y_1}{\partial \theta_1}(t_0) & \cdots & \frac{\partial y_1}{\partial \theta_{p+n}}(t_0) \\ \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial \theta_1}(t_0) & \cdots & \frac{\partial y_m}{\partial \theta_{p+n}}(t_0) \\ \vdots & & \vdots \\ \frac{\partial y_1}{\partial \theta_1}(t_N) & \cdots & \frac{\partial y_1}{\partial \theta_{p+n}}(t_N) \\ \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial \theta_1}(t_N) & \cdots & \frac{\partial y_m}{\partial \theta_{p+n}}(t_N) \end{pmatrix}. \tag{3.6}$$

Our numerical method tests for local structural identifiability since it requires a known set of nominal parameter and initial values at which to compute these sensitivities. A full ranked matrix $\boldsymbol{S}$, is a sufficient condition for local structural identifiability [16, 17]. Rank deficiency of $\boldsymbol{S}$ can be attributed to 2 factors: 1) a model output may be insensitive to a

specific parameter and in this instance, the parameter is classified as unidentifiable and 2) a model output may be sensitive to a particular parameter, but this sensitivity is counteracted by the sensitivity of the model output to one or more other parameters. This implies that these parameters are totally correlated and unidentifiable [18]. The rank of a sensitivity matrix is numerically determined using an SVD:

$$\boldsymbol{S} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T. \tag{3.7}$$

If $\boldsymbol{S}$ has $p+n$ columns, matrix $\boldsymbol{\Sigma}$ will have $p+n$ singular values on its diagonal and these are arranged in descending order. The rank of $\boldsymbol{S}$ is the number of nonzero singular values and conversely rank-deficiency is indicated by the presence of zero-valued singular values [19]. Due to numerical rounding errors, singular values are rarely exactly zero and so one uses as practical definition: Zero-valued singular values are values that fall beyond a distinct gap in the spectrum of singular values [20]. Once possible unidentifiability based on the presence of zero-valued singular values has been established, unidentifiable parameters are recognised as the nonzero entries in the columns of the matrix $\boldsymbol{V}$, related to these vanishing singular values. Both the singular values and the unidentifiable parameters can graphically be illustrated in an easy to interpret identifiability signature [14].

As an example, consider the identifiability signature of the JAK/STAT model in figures 3.1 and 3.2. Defined in section 3.3, the model contains 23 unknown parameters and so there are 23 singular values. Seen in figure 3.1, the 2 singular values beyond the gap suggest that the model is rank deficient and that there are 2 sets of totally correlated parameters. Figure 3.2 shows the elements in the last 2 columns of the $\boldsymbol{V}$ matrix. These corresponds to the 2 singular values beyond the gap in figure 3.1. The nonzero entries in figure 3.2 suggest that parameters $\theta_{11}, \theta_{15}, \theta_{17}, \theta_{21}$ and θ_{22} are structurally unidentifiable.

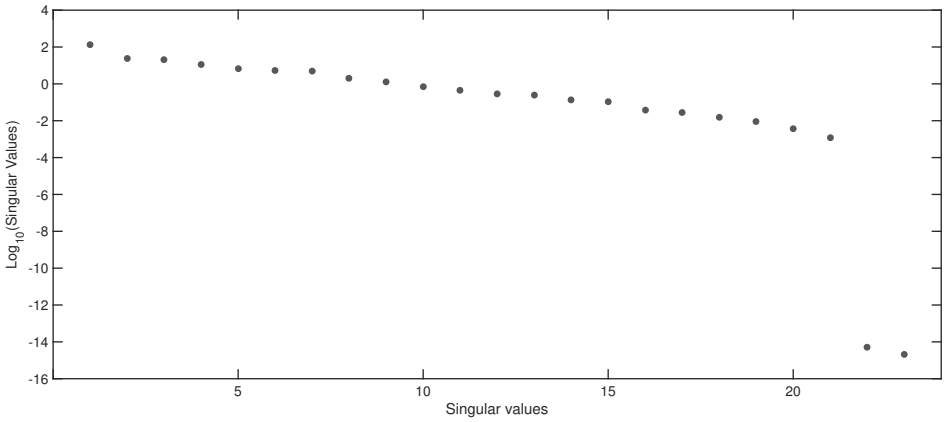


Figure 3.1: **Singular values of the JAK/STAT model** - The 2 singular values beyond the gap suggest that the model is unidentifiable, with these numerically zero-valued singular values alluding to rank deficiency of the sensitivity matrix. These also suggest that there are 2 sets of totally correlated parameters.

To determine which subsets of parameters are totally correlated, we remove a column related to one of the unidentifiable parameters from $\boldsymbol{S}$ and repeat the SVD analysis.

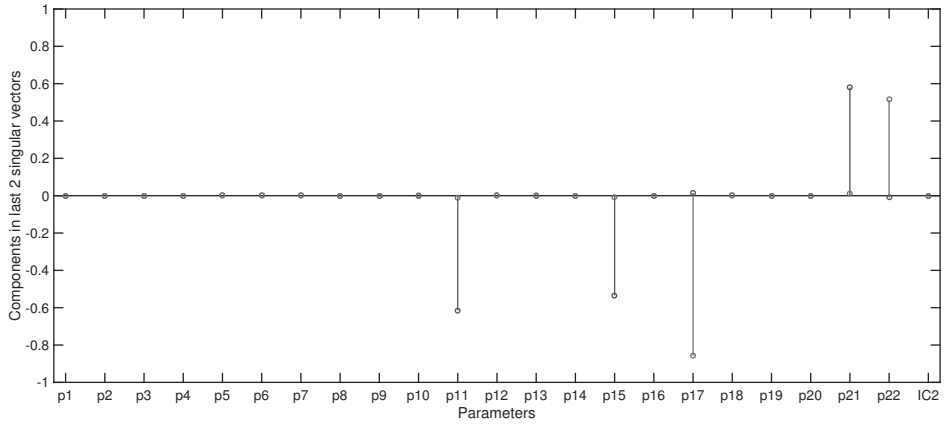


Figure 3.2: **Last 2 columns of the right singular matrix of the JAK/STAT model** - These columns are related to the 2 singular values beyond the gap in figure 3.1. The nonzero elements indicate that parameters $\theta_{11}, \theta_{15}, \theta_{17}, \theta_{21}$ and θ_{22} might be unidentifiable.

This destroys the correlation between parameters in a specific subset and so, only 1 totally correlated set remains. Figure 3.3 shows the last column of the right singular vector, related to the only remaining singular value beyond the gap when column 11, related to parameter θ_{11}, is omitted. The nonzero entries of this column reveal that parameters θ_{17} and θ_{22} remain unidentifiable and therefore are totally correlated. The remaining subset thus contains totally correlated parameters θ_{11}, θ_{15} and θ_{21}. These results can be now be used in the symbolic calculations to: 1) verify the unidentifiability results obtained in step 1, and 2) obtain suggestions for redefined parameters that can be used to reparameterise this model.

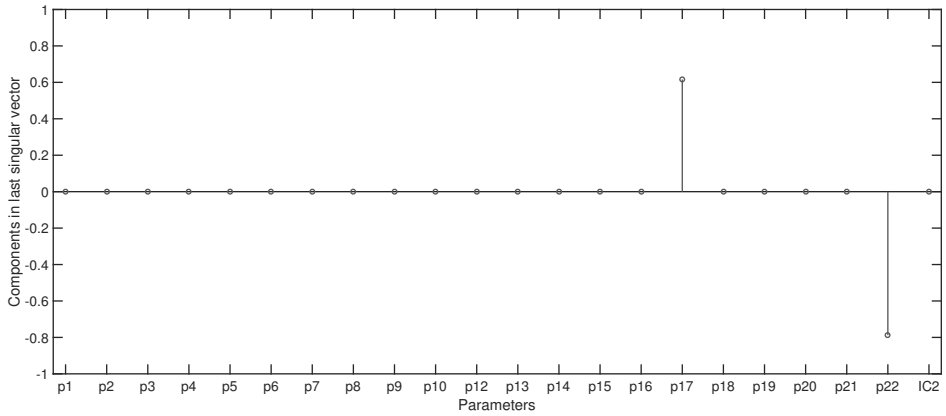


Figure 3.3: **Last column of the right singular matrix of the JAK/STAT model, calculated after the column related to parameter θ_{11} was removed from S.** - The nonzero elements reveal that parameters θ_{17} and θ_{22} belong to a totally correlated or aliased set.

STEP 2: SYMBOLIC IDENTIFIABILITY CALCULATIONS

Symbolic calculations can be preformed using information obtained in step 1. The theoretical characterisation of the identifiability problem for nonlinear systems was already described in the 1970s [21]. The work of Fliess [22] and Tunali and Tarn [23] further made it possible to analyse nonlinear systems symbolically. The Jacobi matrix of a model as defined in (3.1)-(3.3) can be computed using Lie derivatives, where a Lie derivative, $\mathscr{L}_{\boldsymbol{f}}\boldsymbol{h}(\boldsymbol{x})$, is the directional derivative of the smooth function, $\boldsymbol{h}(\boldsymbol{x})$, with respect to the vector field, $\boldsymbol{f}(\boldsymbol{x})$, which describes the model dynamics. A Lie derivative is defined as [24]:

$$\mathscr{L}_{\boldsymbol{f}}\boldsymbol{h}(\boldsymbol{x}) = \frac{\partial \boldsymbol{h}(\boldsymbol{x})}{\partial \boldsymbol{x}}\boldsymbol{f}(\boldsymbol{x}), \tag{3.8}$$

with higher order derivatives computed consecutively as:

$$\mathscr{L}^i_{\boldsymbol{f}}\boldsymbol{h}(\boldsymbol{x}) = \frac{\partial \mathscr{L}^{i-1}_{\boldsymbol{f}}\boldsymbol{h}(\boldsymbol{x})}{\partial \boldsymbol{x}}\boldsymbol{f}(\boldsymbol{x}). \tag{3.9}$$

By parameterising the unknown initial conditions and therefore regarding them as additional parameters, the Jacobi matrix may have up to $p+n$ columns [25]. The augmented parameter vector is defined as, $\boldsymbol{\theta} = \binom{\boldsymbol{\theta}}{\boldsymbol{x_0}}$, and the Jacobi matrix is given by:

$$\frac{\partial \boldsymbol{G}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial \boldsymbol{h}}{\partial \theta_1} & \cdots & \frac{\partial \boldsymbol{h}}{\partial \theta_{p+n}} \\ \frac{\partial \mathscr{L}_{\boldsymbol{f}}\boldsymbol{h}}{\partial \theta_1} & \cdots & \frac{\partial \mathscr{L}_{\boldsymbol{f}}\boldsymbol{h}}{\partial \theta_{p+n}} \\ \frac{\partial \mathscr{L}^2_{\boldsymbol{f}}\boldsymbol{h}}{\partial \theta_1} & \cdots & \frac{\partial \mathscr{L}^2_{\boldsymbol{f}}\boldsymbol{h}}{\partial \theta_{p+n}} \\ \vdots & \cdots & \vdots \end{pmatrix}. \tag{3.10}$$

For identifiability, it is sufficient for $\frac{\partial \boldsymbol{G}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta})$ to have rank $p+n$, implying that all initial values and system parameters can uniquely be determined [26]. It is known from linear algebra that rank deficiency of a matrix is equivalent to it having a nontrivial null-space [27]. Like Anguelova and co-authors, we directly calculate the null-space of the Jacobi matrix [3]. The elements in such a nontrivial null-space reveal the nature of the correlation between the individual unknown parameters. What makes our approach attractive, is that the Jacobi matrix is computed using the preceding numerical results, and so one only has to compute derivatives of the Lie derivatives with respect to the parameters that were suggested to be unidentifiable in step 1. This significantly reduces the computational demand typically associated with symbolic identifiability analyses.

As an example, consider the unidentifiable JAK/STAT model, alluded to in the previous section. The Jacobi matrix now only has 5 columns (instead of 23), each pertaining to an unidentifiable parameter in figure 3.2. The symbolically calculated nontrivial null-space confirms the results in figures 3.1 and 3.2. Using set $\{\theta_{11},\theta_{15},\theta_{17},\theta_{21},\theta_{22}\}$ as possible unidentifiable parameters, we find 2 sets of totally correlated parameters, one

of which is show in figure 3.3. These are verified by the 2 base vectors spanning the null-space: $\mathcal{N}\left(\frac{d\boldsymbol{G}}{d\boldsymbol{\theta}^{\boldsymbol{unid}}}(\boldsymbol{\theta})\right) = \{0,0,-\theta_{17}/\theta_{22},0,1\}\{-\theta_{11}/\theta_{21},-\theta_{15}/\theta_{21},0,1,0\}$, where $\boldsymbol{\theta}^{\boldsymbol{unid}} = \{\theta_{11}, \theta_{15},\theta_{17},\theta_{21},\theta_{22}\}$.

The correlation within these 2 sets of parameters is described by 2 partial differential equations (PDEs) for the variables $\phi_1 = \phi_1(\theta_{17},\theta_{22})$ and $\phi_2 = \phi_2(\theta_{11},\theta_{15},\theta_{21})$:

$$-\frac{\theta_{17}}{\theta_{22}}\frac{\partial\phi_1}{\partial\theta_{17}} + \frac{\partial\phi_1}{\partial\theta_{22}} = 0, \tag{3.11}$$

$$-\frac{\theta_{11}}{\theta_{21}}\frac{\partial\phi_2}{\partial\theta_{11}} - \frac{\theta_{15}}{\theta_{21}}\frac{\partial\phi_2}{\partial\theta_{15}} + \frac{\partial\phi_2}{\partial\theta_{21}} = 0. \tag{3.12}$$

Symbolically obtained solutions to these equations offer suggestions for reparameterisations that will ensure this model's local structural identifiability [12]. Since no unique solutions to these equations exist, solutions can be expressed in terms of different parameters and consequently, biologically relevant reparameterisations should be identified.

Steps 3 and 4 of our reparameterisation schedule will be explained in the context of individual examples in the next section. The reparameterised JAK/STAT model contains 21 parameters in stead of 23. This is related to the fact that there are 2 sets of totally correlated parameters.

3.3. RESULTS

EXAMPLE 1: IMMUNOLOGICAL MODEL FOR MASTITIS IN DAIRY COWS (2 STATES, 7 PARAMETERS)

We first consider a small model to show how the transformation of a model state ensures local structural identifiability. We show that because the initial conditions of states can be regarded as additional parameters, $\boldsymbol{\theta} = \binom{\boldsymbol{\theta}}{\boldsymbol{x_0}}$, the required state transformation is implicitly included in the results. This nonlinear immunological model, describing mastitis in dairy cows, has 2 states and 5 system parameters [28, 29]:

$$\dot{x_1} = \theta_1 x_1 - \theta_2 x_1 x_2, \tag{3.13}$$

$$\dot{x_2} = \theta_3 x_2(1-\theta_4 x_2) + \theta_5 x_1 x_2. \tag{3.14}$$

If the initial conditions of the model states, $x_1(0)$ and $x_2(0)$, are unknown, the model has 7 unknown parameters. This model was found to be unidentifiable measuring the defined model output, $y = \{x_1\}$, with parameters θ_2,θ_4 and the initial condition $x_2(0)$ classified as both unknown and totally correlated [6, 29]. We apply the method outlined in Section 2 and take the following steps: ***Step 1:*** The calculated singular values of this model are shown in figure 3.4. The singular value beyond the gap suggests that the null-space of the sensitivity matrix contains 1 base vector and consequently there may only be 1 set of totally correlated parameters. Figure 3.5 reveals the elements of this set as $\boldsymbol{\theta}^{unid} = \{\theta_2,\theta_4,x_2(0)\}$. A reparameterised model will therefore have 6 unknown parameters instead of 7.

To check whether these findings are correct and find plausible new parameters, we perform ***Step 2:*** The computed Jacobi matrix now has 3 columns only, corresponding to

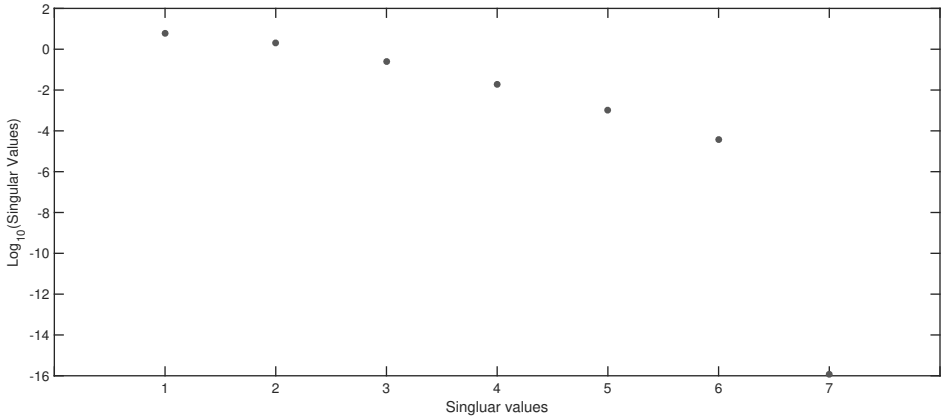


Figure 3.4: **Identifiability signature for the Immunological model** - One numerically zero-valued singular value suggests the rank deficiency of the sensitivity matrix. This singular value falls beyond a gap larger than 3 decades and indicates that there is 1 set of totally correlated parameters.

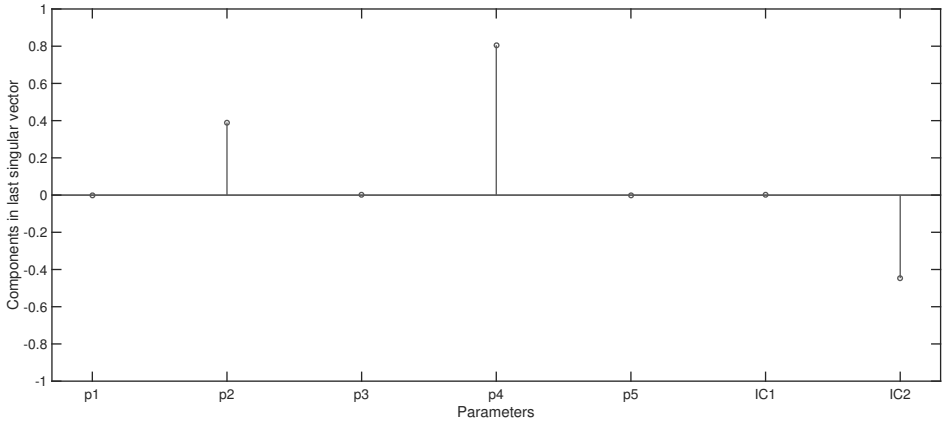


Figure 3.5: **Components of the last column of the right singular vector corresponding to the smallest singular value in figure 3.4** - The nonzero elements in this column indicate that parameters θ_2, θ_4 and initial condition $x_2(0)$ may be totally correlated and consequently unidentifiable.

the 3 unidentifiable parameters. The nontrivial null-space of the Jacobi matrix confirms the numerical findings in step 1. This null-space is spanned by the vector $\mathcal{N}\left(\frac{d\boldsymbol{G}}{d\boldsymbol{\theta}^{\boldsymbol{unid}}}(\boldsymbol{\theta})\right) = \{-\theta_2/x_2(0), -\theta_4/x_2(0), 1\}$. The linear dependence between the relevant columns of the Jacobi matrix is described by the following PDE for some function $\phi\{\theta_2, \theta_4, x_2(0)\}$:

$$-\frac{\theta_2}{x_2(0)}\frac{\partial\phi}{\partial\theta_2} - \frac{\theta_4}{x_2(0)}\frac{\partial\phi}{\partial\theta_4} + \frac{\partial\phi}{\partial x_2(0)} = 0, \tag{3.15}$$

where the coefficients are the elements in the calculated base vector of the null-space. Possible solutions to this partial differential equation are $\phi_1 = \theta_4/\theta_2$ and $\phi_2 = \theta_2 x_2(0)$.

Since we now know that this model is structurally unidentifiable, we may pause to think about the different options we have at our disposal to reinstate this model's identifiability. Recall, that the aim is to destroy the total correlation between the correlated parameters. First, the model will become identifiable if the output set is expanded to include the direct measurement of one of the parameters θ_2, θ_4 or $x_2(0)$. Alternatively, the value of one of these unknown parameters can be assumed as known. If none of these options are viable, we may need to reparameterise the model. ***Step 3:*** In this example, we choose to eliminate system parameter θ_2 from the model. Substituting $\theta_4 = \phi_1\theta_2$ and $x_2(0) = \phi_2/\theta_2$ into the original model, we can identify the state transformation required to eliminate θ_2 from the model as $\tilde{x}_2 = \theta_2 x_2$. Notice, that the state transformation is apparent from our identifiability analysis and accordingly we do not require any additional calculations to obtain this transformation. Consider the model equations, with the substituted parameters highlighted for convenience:

$$\dot{x_1} = \theta_1 x_1 - \theta_2 x_2 x_1, \tag{3.16}$$

$$\dot{x_2} = \theta_3 x_2 (1 - \boldsymbol{\phi_1 \theta_2} x_2) + \theta_5 x_1 x_2. \tag{3.17}$$

The initial conditions should now be taken as $x_1(0)$ and $x_2(0) = \boldsymbol{\phi_2/\theta_2}$ respectively. Multiplying equation 3.17 by θ_2, yields the final form of the reparameterised identifiable model, now with 6 parameters instead of 7 (θ_1, θ_3, θ_5, ϕ_1, $x_1(0)$, ϕ_2):

$$\dot{x_1} = \theta_1 x_1 - x_1 \tilde{x}_2, \tag{3.18}$$

$$\dot{\tilde{x}}_2 = \theta_3 \tilde{x}_2 (1 - \phi_1 \tilde{x}_2) + \theta_5 x_1 \tilde{x}_2, \tag{3.19}$$

with $x_1(0)$ and $\tilde{x}_2(0) = \phi_2$. In the supplementary material, we show that the original and reparameterised models have the same solutions and that the reparameterised model is structurally identifiable.

EXAMPLE 2: MICROBIAL GROWTH MODEL (2 STATES, 6 PARAMETERS)

In this example we consider a benchmark model that also requires a state transformation to obtain an identifiable reparametrisation [9]. Here, we illustrate how an unidentifiable initial condition of one of the model states is eliminated. This model describes the microbial growth in a batch reactor and has 2 states and 6 parameters [30]:

$$\dot{x} = \frac{\mu x s}{K_s + s} - K_d x, \tag{3.20}$$

$$\dot{s} = -\frac{\mu x s}{Y(K_s + s)}, \tag{3.21}$$

with unknown initial conditions $x(0)$ and $s(0)$. This model is unidentifiable measuring the output $\boldsymbol{y} = \{x\}$, with parameters K_s, Y and the initial condition of state $s(0)$ not estimable [9]. ***Step 1:*** The calculated singular values seen in figure 3.6, show a single singular value beyond the gap. This suggests that the null-space of the sensitivity matrix contains only one base vector and so all unidentifiable parameters are also totally correlated. The parameters in this set, $\boldsymbol{\theta}^{unid} = \{K_s, Y, s(0)\}$, can be deduced from figure 3.7.

To confirm these results, we continue with ***Step 2:*** Using the results from step 1, the calculated Jacobi matrix now only has 3 columns. The nontrivial null-space of this matrix is calculated as $\mathcal{N}\left(\frac{d\boldsymbol{G}}{d\boldsymbol{\theta}^{unid}}(\boldsymbol{\theta})\right) = \{K_s/s(0), -Y/s(0), 1\}$, and so the number of model

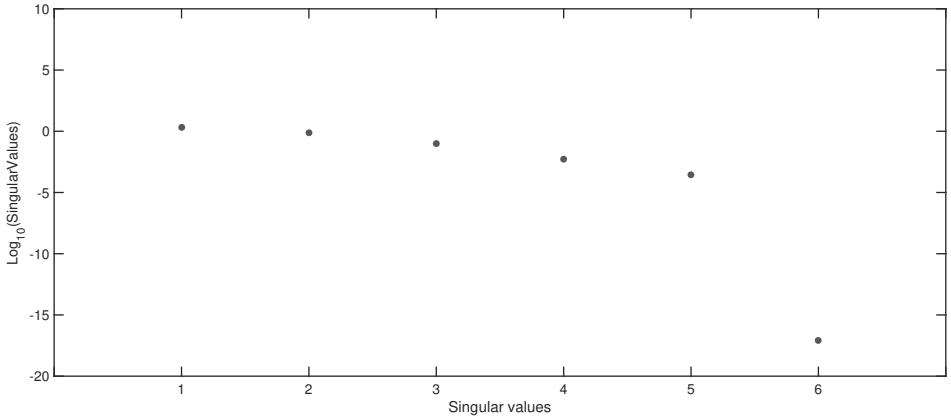


Figure 3.6: **Identifiability signature for the Microbial growth model** - One zero-valued singular value suggests the rank deficiency of the sensitivity matrix. This singular value falls beyond a gap larger than 3 decades and suggests that there is 1 set with totally correlated parameters.

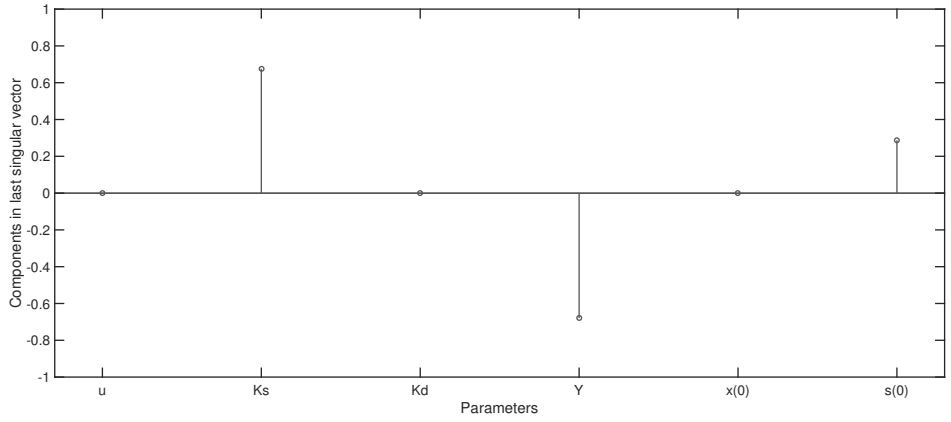


Figure 3.7: **Components of the last column of the right singular vector corresponding to the smallest singular value in figure 3.6** - The nonzero elements in this column indicate that parameters K_s, Y and initial condition $s(0)$ might be totally correlated and therefore unidentifiable.

parameters has to be reduced from 6 to 5. This leads to the following PDE for some function $\phi\{K_s, Y, s(0)\}$:

$$\frac{K_s}{s(0)}\frac{\partial\phi}{\partial K_s} - \frac{Y}{s(0)}\frac{\partial\phi}{\partial Y} + \frac{\partial\phi}{\partial s(0)} = 0. \tag{3.22}$$

Possible solutions to this equation are $\phi_1 = K_s/s(0)$, and $\phi_2 = Y s(0)$. This model could be made structurally identifiability if its measured output is expanded to include one of the unidentifiable parameters, for example $\boldsymbol{y} = \{x, s\}$, or by assuming the value of one of these unknowns as known. Here, we choose to reparameterise the model. Choos-

ing to eliminate the unidentifiable initial condition, $s(0)$, from the model requires the substitution of parameters K_s and Y. After rearranging the solutions to 3.22, one obtains suitable substitutions for these parameters, $K_s = \phi_1 s(0)$ and $Y = \phi_2 / s(0)$. ***Step 3:*** After substitution and multiplying 3.21 by $1/s(0)$, some re-arranging reveals the state transformation required to ensure the model's structural identifiability as $\tilde{s} = s/s(0)$:

$$\dot{x} = \frac{\mu x s}{\boldsymbol{\phi_1 s(0)} + s} - K_d x = \frac{\mu x_1 \frac{s}{s(0)}}{\phi_1 + \frac{s}{s(0)}} - K_d x, \tag{3.23}$$

$$\frac{1}{s(0)} \dot{s} = -\frac{1}{s(0)} \frac{\mu x s}{\frac{\boldsymbol{\phi_2}}{\boldsymbol{s(0)}} (\boldsymbol{\phi_1 s(0)} + s)} = -\frac{\mu x \frac{s}{s(0)}}{\phi_2 (\phi_1 + \frac{s}{s(0)})}. \tag{3.24}$$

The final model contains 5 parameters instead of 6 (μ, K_d, ϕ_1, ϕ_2 and $x(0)$). The initial condition of s is now known since $\tilde{s}(0) = 1$. We arrive at the reparameterisation that has been reported by Evans and Chappell [9].

$$\dot{x} = \frac{\mu x \tilde{s}}{\phi_1 + \tilde{s}} - K_d x, \tag{3.25}$$

$$\dot{\tilde{s}} = -\frac{\mu x \tilde{s}}{\phi_2 (\phi_1 + \tilde{s})}. \tag{3.26}$$

In the supplementary material we show that the original and new models have exactly the same solutions.

It might be the case that the system defined in 3.25 and 3.26 is not biologically relevant, necessitating an alternative reparameterisation. Substituting $K_s = \phi_1 s(0)$ and $Y = \phi_2 / s(0)$ into equations 3.20 and 3.21 and multiplying the both equations 3.27 and 3.28 by Y, reveals a state transformation that might be biologically relevant as $\tilde{s} = Y s$:

$$\dot{x} = (\frac{Y}{Y}) \frac{\mu x s}{\boldsymbol{\phi_1 s(0)} + s} - K_d x = \frac{\mu x Y s}{\frac{\phi_2}{\cancel{s(0)}} \phi_1 \cancel{s(0)} + Y s} - K_d x, \tag{3.27}$$

$$Y \dot{s} = -Y \frac{\mu x s}{(Y K_s + Y s)} = -\frac{\mu x Y s}{(\frac{\phi_2}{\cancel{s(0)}} \phi_1 \cancel{s(0)} + Y s)}. \tag{3.28}$$

The reparameterised model has 5 unknown parameters (μ, K_d, ϕ_1, ϕ_2 and $x(0)$) and reduces to:

$$\dot{x} = \frac{\mu x \tilde{s}}{\phi_2 \phi_1 + \tilde{s}} - K_d x, \tag{3.29}$$

$$\dot{\tilde{s}} = -\frac{\mu x 1 \tilde{s}}{\phi_2 \phi_1 + \tilde{s}}, \tag{3.30}$$

with initial conditions now defied as $x(0)$ and $\tilde{s}(0) = \phi_2$ respectively.

EXAMPLE 3: JAK/STAT MODEL (14 STATES, 23 PARAMETERS)

Let us now consider a large model, for which the reparameterisation process requires 2 state transformations. The constitutive activation of the JAK (Janus kinase)/STAT signalling pathway forms part of both the primary mediastinal B-cell lymphoma (PMBL) and the classical Hodgkin lymphoma (cHL) [31]. Raue *et al.* investigated the identifiability of this benchmark model using three different approaches and concluded that the model is unidentifiable [32]. The initial value of state x_2 is unknown and regarded as an additional parameter and so in total, 23 parameters need to be inferred [32, 33]:

$$\dot{x}_1 = -\theta_1 u_1 c_1 x_1 - \theta_5 x_1 + \theta_6 x_2, \tag{3.31}$$
$$\dot{x}_2 = \theta_5 x_1 - \theta_6 x_2, \tag{3.32}$$
$$\dot{x}_3 = \theta_1 u_1 c_1 x_1 - \theta_2 x_3 x_7, \tag{3.33}$$
$$\dot{x}_4 = \theta_2 x_3 x_7 - \theta_3 x_4, \tag{3.34}$$
$$\dot{x}_5 = \theta_3 x_4 - \theta_4 x_5, \tag{3.35}$$
$$\dot{x}_6 = -\frac{\theta_7 x_3 x_6}{(1+\theta_{13} x_{13})} - \frac{\theta_7 x_4 x_6}{(1+\theta_{13} x_{13})} + \theta_8 c_2 x_7, \tag{3.36}$$
$$\dot{x}_7 = \frac{\theta_7 x_3 x_6}{(1+\theta_{13} x_{13})} + \frac{\theta_7 x_4 x_6}{(1+\theta_{13} x_{13})} - \theta_8 c_2 x_7, \tag{3.37}$$
$$\dot{x}_8 = -\theta_9 x_8 x_7 + c_2 \theta_{10} x_9, \tag{3.38}$$
$$\dot{x}_9 = \theta_9 x_8 x_7 - c_2 \theta_{10} x_9, \tag{3.39}$$
$$\dot{x}_{10} = \theta_{11} x_9, \tag{3.40}$$
$$\dot{x}_{11} = -\theta_{12} c_1 u_1 x_{11}, \tag{3.41}$$
$$\dot{x}_{12} = \theta_{12} c_1 u_1 x_{11}, \tag{3.42}$$
$$\dot{x}_{13} = \frac{\theta_{14} x_{10}}{(\theta_{15} + x_{10})} - \theta_{16} x_{13}, \tag{3.43}$$
$$\dot{x}_{14} = \theta_{17} x_9. \tag{3.44}$$

The model output contains 5 additional parameters:

$$y_1 = x_1 + x_3 + x_4, \tag{3.45}$$
$$y_2 = \theta_{18}(x_3 + x_4 + x_5 + x_{12}), \tag{3.46}$$
$$y_3 = \theta_{19}(x_4 + x_5), \tag{3.47}$$
$$y_4 = \theta_{20} x_7, \tag{3.48}$$
$$y_5 = \theta_{21} x_{10}, \tag{3.49}$$
$$y_6 = \theta_{22} x_{14}, \tag{3.50}$$
$$y_7 = x_{13}, \tag{3.51}$$
$$y_8 = x_9. \tag{3.52}$$

The initial values of the individual model states are $\boldsymbol{x}(0) = \{1.3, \theta_{23}, 0, 0, 0, 2.8, 0, 165, 0, 0, 0.34, 0, 0, 0\}$ [32]. The constants c_1, c_2 and model input u_1 are known. ***Step 1:*** Some initial conditions are defined as zero, so we calculate the sensitivity matrix $\boldsymbol{S}$. The model's unidentifiability is evident from the large gap between the singular values seen

in figure 3.1. The 2 singular values beyond the gap suggest that the null-space contains 2 base vectors and so there are 2 sets of totally correlated parameters. The union of the elements in these 2 sets, $\boldsymbol{\theta}^{unid} = \{\theta_{11}, \theta_{15}, \theta_{17}, \theta_{21}, \theta_{22}\}$, can be identified as the nonzero elements in figure 3.2. This is in agreement with the findings of previous authors [32]. To determine which parameters are totally correlated, a column of $\boldsymbol{S}$, related to an unidentifiable parameter, is omitted. Figure 3.3 reveals that parameters θ_{17} and θ_{22} remain unidentifiable when the column related to parameter θ_{11} is removed and so we conclude that they belong to the same totally correlated set. The remaining parameters, θ_{11}, θ_{15} and θ_{21}, apparently form the second set of totally correlated parameters. The total number of parameters will therefore have to be reduced from 23 to 21.

Step 2: The calculated Jacobi matrix now only has 5 columns, each corresponding to an unidentifiable parameter in figure 3.2. The computed nontrivial null-space of this Jacobian confirms the findings in step 1. This null-space $\mathcal{N}\left(\frac{d\boldsymbol{G}}{d\boldsymbol{\theta}^{\boldsymbol{unid}}}(\boldsymbol{\theta})\right)$ is spanned by the vectors $\{0, 0, -\theta_{17}/\theta_{22}, 0, 1\}$,$\{-\theta_{11}/\theta_{21}, -\theta_{15}/\theta_{21}, 0, 1, 0\}$. The corresponding PDEs are given in equations 3.11 and 3.12 respectively. Possible solutions to these equations include: $\phi_{1,1} = \theta_{17}\theta_{22}$ as a solution to 3.11, and $\phi_{2,1} = \theta_{15}/\theta_{11}$ and $\phi_{2,2} = \theta_{11}\theta_{21}$ as solutions to 3.12.

Before reparameterising this model, let us first reflect on the different options available to obtain its structural identifiability. These include: 1) adding sensors to the model's output, thereby addressing the topic of minimal outputs sets. For example, if parameters θ_{17} and θ_{21} were to be added as measured sensors, the model would become identifiable. 2) alternatively, one can explore the possibility of assuming the value of one of the unidentifiable parameters in each set as known. 3) interestingly, this model's unidentifiability can also be attributed to its predefined initial conditions [5]. All 3 these options may however be unrealistic and in such cases, reparameterisation is needed.

Step 3: The mentioned solutions to 3.11 and 3.12 lead to the following possible parametric substitutions: $\theta_{22} = \phi_{1,1}/\theta_{17}, \theta_{15} = \theta_{11}\phi_{2,1}$ and $\theta_{21} = \phi_{2,2}/\theta_{11}$. Choosing to eliminate θ_{11} and θ_{17} reveals the 2 required state transformations as $\tilde{x}_{10} = x_{10}/\theta_{11}$ and $\tilde{x}_{14} = x_{14}/\theta_{17}$ respectively. The substituted parameters are highlighted for clarity:

$$\frac{\dot{x}_{10}}{\theta_{11}} = x_9, \tag{3.53}$$

$$\dot{x}_{13} = \frac{\theta_{14}x_{10}}{\boldsymbol{\theta_{11}\phi_{2,1}} + x_{10}} - \theta_{16}x_{13} = \frac{\theta_{14}x_{10}}{\theta_{11}(\phi_{2,1} + \frac{x_{10}}{\theta_{11}})} - \theta_{16}x_{13}, \tag{3.54}$$

$$\frac{\dot{x}_{14}}{\theta_{17}} = x_9, \tag{3.55}$$

Substituting θ_{21} and θ_{22} into the relevant model outputs, yields:

$$y_5 = \frac{\boldsymbol{\phi_{2,2}}}{\boldsymbol{\theta_{11}}} x_{10}, \tag{3.56}$$

$$y_6 = \frac{\boldsymbol{\phi_{1,1}}}{\boldsymbol{\theta_{17}}} x_{14}. \tag{3.57}$$

We refer to the supplementary material for the final model structure, the accompanying structural identifiability results and conformation that the 2 models have the same

solutions.

EXAMPLE 4: LUNG CANCER MODEL (21 STATES, 75 PARAMETERS)

In this example we apply our method to a large system of ordinary differential equations. In addition, we illustrate the process of model analysis that we think should be followed when analysing a model's identifiability. The model in question, uses a systems biology approach to understand the biology of the Epidermal Growth Factor Receptor (EGFR) and type 1 Insulin-like Growth Factor (IGF1R) pathways in non-small cell lung cancer (NSCLC) [34, 35]. The authors present a detailed in silico ordinary differential equation model, which consists of 21 states and 54 system parameters. In the original paper, the values of all but 5 of these parameters values are assumed known and are taken from literature. To analyse this model's structural identifiability, we assume that the values of all 54 system parameters and the 21 initial conditions of the model states are unknown. We also assume that only certain individual model states are measurable sensors.

A good starting point in the identifiability analysis of a model is to look at its directed graph. A directed graph is a graphical representation of an ODE system and depicts the connectivity between individual states. This gives us a visual cue as to the interaction between model states as well as hinting towards states/sensors that need to be measured to ensure structural identifiability. As models get larger, this visual analysis becomes cumbersome and so one has to proceed with a more theoretical approach. We know that sensors x_1 and x_9, related to EGFRactive and IGFRactive respectively, are both root strongly connected components and so need to be measured to ensure the model's structural identifiability. To illustrate our reparameterisation procedure, we will assume that these key states cannot be measured and so we analyse the model measuring as output, $\boldsymbol{y} = [x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{17}, x_{18}, x_{19}, x_{20}, x_{21}]$.

Step 1: The model is structurally unidentifiable measuring $\boldsymbol{y}$. In figure 3.8 we observe 2 singular values beyond the gap, suggesting that the null-space spans 2 base vectors and so there are 2 sets of totally correlated parameters. Furthermore, this tells us that if we wish to reparameterise this model, the total number of unknown parameters will have to be reduced by 2. The unidentifiable parameters, $\boldsymbol{\theta}^{unid} = \{\theta_6, \theta_{17}, \theta_{20}, \theta_{23}, x_1(0), x_9(0)\}$, can be deducted from figure 3.9. The full model is given in the supplementary material. Here, we only give the equations containing the possibly unidentifiable parameters:

$$\dot{x}_1 = -\theta_5 x_1, \tag{3.58}$$

$$\dot{x}_2 = -\frac{\theta_6 x_1 x_2^{\theta_7}}{\theta_8^{\theta_7} - x_2^{\theta_7}} + \frac{x_{21}\theta_{14} x_3^{\theta_{15}}}{\theta_{16}^{\theta_{15}} + x_3^{\theta_{15}}} - \frac{\theta_{17} x_9 x_2^{\theta_{18}}}{\theta_{19}^{\theta_{18}} + x_2^{\theta_{18}}}, \tag{3.59}$$

$$\dot{x}_3 = \frac{\theta_6 x_1 x_2^{\theta_7}}{\theta_8^{\theta_7} - x_2^{\theta_7}} - \frac{x_{21}\theta_{14} x_3^{\theta_{15}}}{\theta_{16}^{\theta_{15}} + x_3^{\theta_{15}}} + \frac{\theta_{17} x_9 x_2^{\theta_{18}}}{\theta_{19}^{\theta_{18}} + x_2^{\theta_{18}}}, \tag{3.60}$$

$$\dot{x}_9 = -\theta_1 x_9, \tag{3.61}$$

$$\dot{x}_{10} = -\frac{x_9\theta_{20} x_{10}^{\theta_{21}}}{\theta_{22}^{\theta_{21}} + x_{10}^{\theta_{21}}} - \frac{x_1^2\theta_{23} x_{10}^{\theta_{24}}}{\theta_{25}^{\theta_{24}} + x_{10}^{\theta_{24}}} - \frac{x_5\theta_{33} x_{10}^{\theta_{34}}}{\theta_{35}^{\theta_{34}} + x_{10}^{\theta_{34}}} + \theta_2 x_{11}, \tag{3.62}$$

$$\dot{x}_{11} = \frac{x_9\theta_{20}x_{10}^{\theta_{21}}}{\theta_{22}^{\theta_{21}} + x_{10}^{\theta_{21}}} + \frac{x_1^2\theta_{23}x_{10}^{\theta_{24}}}{\theta_{25}^{\theta_{24}} + x_{10}^{\theta_{24}}} + \frac{x_5\theta_{33}x_{10}^{\theta_{34}}}{\theta_{35}^{\theta_{34}} + x_{10}^{\theta_{34}}} - \theta_2 x_{11}. \tag{3.63}$$

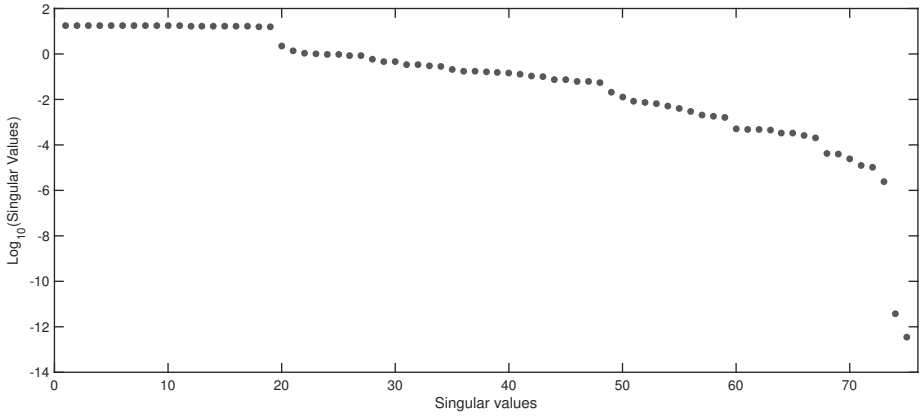


Figure 3.8: **Identifiability signature for the Lung cancer model** - Two vanishing singular values suggest the rank deficiency of the sensitivity matrix and that there may be 2 sets of totally correlated parameters.

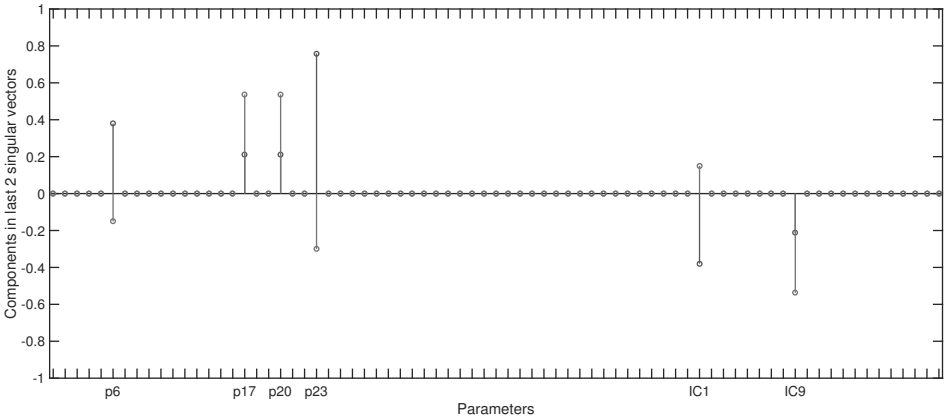


Figure 3.9: **Components of the 2 last columns of the right singular vector corresponding to the 2 vanishing singular values in figure 3.8** - The nonzero elements in this column suggest that parameters $\theta_6, \theta_{17}, \theta_{20}, \theta_{23}$ and initial conditions $x_1(0)$ and $x_9(0)$ are structurally unidentifiable.

As seen before, we can determine which sets of unidentifiable parameters are totally correlated by removing one of the columns from the sensitivity matrix, related to an unidentifiable parameter. For example, figure 3.10 reveals that parameters θ_{17}, θ_{20} and $x_9(0)$ remain unidentifiable when we remove the column related to θ_6 and so we can conclude that these parameters are totally correlated. The remaining parameters, θ_6, θ_{23} and $x_1(0)$, belong to a second set of totally correlated parameters.

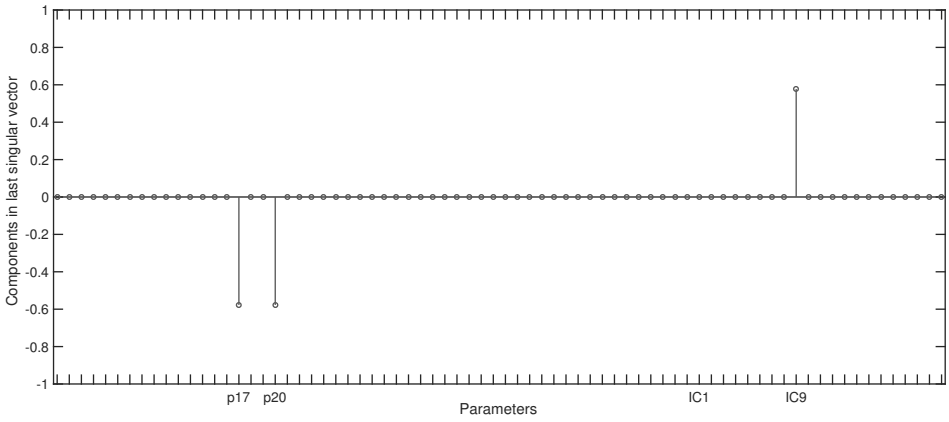


Figure 3.10: **Components of the last column of the right singular vector corresponding to one vanishing singular value.** - If we remove the column of the sensitivity matrix pertaining to parameter θ_6 and re-evaluate the SVD, the nonzero elements in this column suggest that parameters θ_{17}, θ_{20} and initial condition $x_9(0)$ are totally correlated.

Step 2: The Jacobi matrix now only contains 6 columns, each related to an unidentifiable parameter identified in step 1. If we were to omit step 1, we would have to calculate this matrix with 75 columns instead. As expected, the symbolically calculated nullspace of the Jacobian is spanned by 2 base vectors, confirming our findings in step 1: $\mathcal{N}\left(\frac{d\boldsymbol{G}}{d\boldsymbol{\theta}^{\boldsymbol{unid}}}(\boldsymbol{\theta})\right) = \{0, -\theta_{17}/x_9(0), -\theta_{20}/x_9(0), 0, 0, 1\}\{-\theta_6/x_1(0), 0, 0, -2\theta_{23}/x_1(0), 1, 0\}$. The correlation between these parameters is described by the 2 PDEs for the variables $\phi_1 = \phi_1(\theta_{17}, \theta_{20}, x_9(0))$ and $\phi_2 = \phi_2(\theta_6, \theta_{23}, x_1(0))$:

$$-\frac{\theta_{17}}{x_9(0)}\frac{\partial\phi_1}{\partial\theta_{17}} - \frac{\theta_{20}}{x_9(0)}\frac{\partial\phi_1}{\partial\theta_{20}} + \frac{\partial\phi_1}{\partial x_9(0)} = 0, \tag{3.64}$$

$$-\frac{\theta_6}{x_1(0)}\frac{\partial\phi_2}{\partial\theta_6} - 2\frac{\theta_{23}}{x_1(0)}\frac{\partial\phi_2}{\partial\theta_{23}} + \frac{\partial\phi_2}{\partial x_1(0)} = 0. \tag{3.65}$$

Solutions include: $\phi_{1,1} = \theta_{20}/\theta_{17}$ and $\phi_{1,2} = \theta_{17}x_9(0)$ as solutions to 3.64 and $\phi_{2,1} = \theta_{23}/\theta_6^2$ and $\phi_{2,2} = \theta_6 x_1(0)$ as solutions to 3.65. Available options to address this model's unidentifiability include: 1) assume the values of at least one of the parameters in each of the totally correlated sets as known. For example, if the values of θ_6 and $x_9(0)$ are known, the model will be structurally identifiable. 2) Both the directed graph and a minimal output set analysis reveal that if x_1 and x_9 are also measured, this model will be identifiable. 3) If these options cannot be implemented one will have to reparameterise the model.

Step 3: In this example, we choose to eliminate parameters θ_6 and θ_{17} from the model, but as always, it is up to a modeller's discretion which parameters to elliminate. The unidentifiable parameters are now defined as $\theta_{20} = \phi_{1,1}\theta_{17}, x_9(0) = \phi_{1,2}/\theta_{17}, \theta_{23} = \phi_{2,1}\theta_6^2$ and $x_1(0) = \phi_{2,2}/\theta_6$ and substituted. This reveals the 2 required state transformations as: $\tilde{x}_1 = \theta_6 x_1$ and $\tilde{x}_9 = \theta_{17}x_9$:

$$\boldsymbol{\theta_6}\dot{\boldsymbol{x_1}} = -\theta_5\boldsymbol{\theta_6}\boldsymbol{x_1}, \tag{3.66}$$

$$\dot{x}_2 = -\frac{\boldsymbol{\theta_6}\boldsymbol{x_1}x_2^{\theta_7}}{\theta_8^{\theta_7}-x_2^{\theta_7}} + \frac{x_{21}\theta_{14}x_3^{\theta_{15}}}{\theta_{16}^{\theta_{15}}+x_3^{\theta_{15}}} - \frac{\boldsymbol{\theta_{17}}\boldsymbol{x_9}x_2^{\theta_{18}}}{\theta_{19}^{\theta_{18}}+x_2^{\theta_{18}}}, \tag{3.67}$$

$$\dot{x}_3 = \frac{\boldsymbol{\theta_6}\boldsymbol{x_1}x_2^{\theta_7}}{\theta_8^{\theta_7}-x_2^{\theta_7}} - \frac{x_{21}\theta_{14}x_3^{\theta_{15}}}{\theta_{16}^{\theta_{15}}+x_3^{\theta_{15}}} + \frac{\boldsymbol{\theta_{17}}\boldsymbol{x_9}x_2^{\theta_{18}}}{\theta_{19}^{\theta_{18}}+x_2^{\theta_{18}}}, \tag{3.68}$$

$$\boldsymbol{\theta_{17}}\dot{\boldsymbol{x_9}} = -\theta_1\boldsymbol{\theta_{17}}\boldsymbol{x_9}, \tag{3.69}$$

$$\dot{x}_{10} = -\frac{\boldsymbol{x_9}\boldsymbol{\phi_{1,1}}\boldsymbol{\theta_{17}}x_{10}^{\theta_{21}}}{\theta_{22}^{\theta_{21}}+x_{10}^{\theta_{21}}} - \frac{\boldsymbol{x_1^2}\boldsymbol{\phi_{2,1}}\boldsymbol{\theta_6^2}x_{10}^{\theta_{24}}}{\theta_{25}^{\theta_{24}}+x_{10}^{\theta_{24}}} - \frac{x_5\theta_{33}x_{10}^{\theta_{34}}}{\theta_{35}^{\theta_{34}}+x_{10}^{\theta_{34}}} + \theta_2 x_{11}, \tag{3.70}$$

$$\dot{x}_{11} = \frac{\boldsymbol{x_9}\boldsymbol{\phi_{1,1}}\boldsymbol{\theta_{17}}x_{10}^{\theta_{21}}}{\theta_{22}^{\theta_{21}}+x_{10}^{\theta_{21}}} + \frac{\boldsymbol{x_1^2}\boldsymbol{\phi_{2,1}}\boldsymbol{\theta_6^2}x_{10}^{\theta_{24}}}{\theta_{25}^{\theta_{24}}+x_{10}^{\theta_{24}}} + \frac{x_5\theta_{33}x_{10}^{\theta_{34}}}{\theta_{35}^{\theta_{34}}+x_{10}^{\theta_{34}}} - \theta_2 x_{11}. \tag{3.71}$$

The unknown initial conditions are now defined as $\tilde{x}_1(0) = x_1(0)\theta_6 = \phi_{2,2}$ and $\tilde{x}_9(0) = x_9(0)\theta_{17} = \phi_{1,2}$ respectively. Refer to the supplementary material for the final model structure.

3.4. CONCLUSIONS

In this paper, we introduce an efficient method to reparameterise large structurally unidentifiable models. Our hybrid method starts with a numerical identifiability analysis. This analysis is tasked with spotting possible unidentifiable parameters and detecting both the number of totally correlated sets as well as which parameters belong to each of the individual sets. These numerical results are used in subsequent symbolic calculations. Traditionally, these calculations are computationally demanding, sometimes requiring days to obtain the relevant results. However, applying the numerical results allows us to calculate a much smaller Jacobi matrix. This reduced matrix now only contains columns related to unidentifiable parameters and this enables us to analyse large systems biology models within minutes. Symbolic calculations allow for the verification of numerical results and yield suitable suggestions for new parameters to be used in model reparameterisation. As seen, this reparameterisation process may include the redefinition of certain model states by means of state transformations. For some examples, these transformations may immediately be apparent due to the fact that initial conditions can easily be incorporated into a model's analysis by regarding them as additional parameters.

From the examples in section 3.3, we can conclude that not only does our new approach yield the correct results, since they are identical to outcomes obtained in a classical way, but importantly, its computational efficiency allows for extensions into addressing unidentifiability issues of large ODE models. In doing so, modellers can potentially save time and effort in the model development process.

REFERENCES

[1] D. J. Cole, B. J. T. Morgan, and D. M. Titterington, *Determining the parametric structure of models,* Mathematical Biosciences **228**, 16 (2010).

[2] P. Li and Q. D. Vu, *Identification of parameter correlations for parameter estimation in dynamic biological models,* BMC Systems Biology **7**, 1 (2013).

[3] M. Anguelova, J. Karlsson, and M. Jirstrand, *Minimal output sets for identifiability,* Mathematical Biosciences **239**, 139 (2012).

[4] A. F. Villaverde and J. R. Banga, *Dynamical compensation and structural identifiability of biological models: Analysis, implications, and reconciliation,* PLOS Computational Biology **11**, 1 (2017).

[5] M. P. Saccomani, S. Audoly, and L. D'Angió, *Parameter identifiability of nonlinear systems: the role of initial conditions,* Automatica **39**, 619 (2003).

[6] N. Meshkat, M. Eisenberg, and J. J. DiStefano, *An algorithm for finding globally identifiable parameter combinations of nonlinear ode models using Gröbner Bases,* Mathematical Biosciences **222**, 61 (2009).

[7] N. Meshkat, C. Anderson, and J. J. DiStefano, *Finding identifiable parameter combinations in nonlinear ode models and the rational reparameterization of their input–output equations,* Mathematical Biosciences **233**, 19 (2011).

[8] M. J. Chappell and R. N. Gunn, *A procedure for generating locally identifiable reparameterisations of unidentifiable non-linear systems by the similarity transformation approach,* Mathematical Biosciences **148**, 21 (1998).

[9] N. D. Evans and M. J. Chappell, *Extensions to a procedure for generating locally identifiable reparameterisations of unidentifiable systems,* Mathematical Biosciences **168**, 137 (2000).

[10] M. C. Eisenberg and M. A. Hayashi, *Determining identifiable parameter combinations using subset profiling,* Mathematical Biosciences **256**, 116 (2014).

[11] Y. Chu and J. Hahn, *Parameter set selection via clustering of parameters into pairwise indistinguishable groups of parameters,* Ind. Eng. Chem. Res. **48**, 6000–6009 (2009).

[12] R. Choquet and D. Cole, *A hybrid symbolic-numerical method for determining model structure,* Mathematical Biosciences **236**, 117 (2012).

[13] J. D. Stigter and J. Molenaar, *A fast algorithm to assess local structural identifiability,* Automatica **58**, 118 (2015).

[14] J. D. Stigter, D. Joubert, and J. Molenaar, *Observability of complex systems: Finding the gap,* Scientific Reports **7**, 1 (2017).

[15] J. Reid, *Structural identifiability in linear time invariant systems,* IEEE Transactions on Automatic Control **22**, 242 (1977).

[16] Y. Bard, *Nonlinear Parameter Estimation* (Academic Press Inc, 1974).

[17] M. Vidyasagar, *Nonlinear systems analysis* (Prentice Hall, Englewood Cliffs, NJ, 1993).

[18] A. Gábor, A. F. Villaverde, and J. R. Banga, *Parameter identifiability analysis and visualization in large-scale kinetic models of biosystems,* BMC Systems Biology **11**, 1 (2017).

[19] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th ed. (The Johns Hopkins University Press, 2013).

[20] G. Quintana-Ortí and E. S. Quintana-Ortí, *Parallel codes for computing the numerical rank,* Linear Algebra and its Applications **275-276**, 451 (1998).

[21] R. Hermann and A. Krener, *Nonlinear controllability and observability,* IEEE Transactions on Automatic Control **22**, 728 (1977).

[22] M. Fliess, *Fonctionnelles causales non linéaires et indéterminées non commutatives,* Bull Soc Math France **109**, 3 (1981).

[23] E. T. Tunali and T. J. Tarn, *New results for identifiability of nonlinear systems,* IEEE Transactions on Automatic Control **32**, 146 (1987).

[24] A. F. Villaverde, A. Barreiro, and A. Papachristodoulou, *Structural identifiability of dynamic systems biology models,* PLOS Computational Biology **20**, 1 (2016).

[25] J. D. Stigter and R. L. M. Peeters, *On a geometric approach to the structural identifiability problem and its application in a water quality case study,* in *2007 European Control Conference (ECC)* (2007) pp. 3450–3456.

[26] H. Pohjanpalo, *Systems identifiability based on the power series expansion of the solution,* Mathematical Biosciences **41**, 21 (1978).

[27] J. D. Stigter, M. B. Beck, and J. Molenaar, *Assessing local structural identifiability for environmental models,* Environmental Modelling and Software **93**, 398 (2017).

[28] D. Döpfer, *Recurrent clinical Escherichia coli mastitis in dairy cows*, Ph.D. thesis, Utrecht University (2000).

[29] G. Margaria, E. Riccomagno, M. J. Chappell, and H. P. Wynn, *Differential algebra methods for the study of the structural identifiability of rational function state-space models in the biosciences,* Mathematical Biosciences **174**, 1 (2001).

[30] A. Holmberg and J. Ranta, *Procedures for parameter and state estimation of microbial growth process models,* Automatica **18**, 181 (1982).

[31] V. Raia, M. Schilling, M. Böhm, B. Hahn, A. Kowarsch, A. Raue, C. Sticht, S. Bohl, M. Saile, P. Möller, N. Gretz, J. Timmer, F. Theis, W.-D. Lehmann, P. Lichter, and U. Klingmüller, *Dynamic mathematical modeling of IL13-induced signaling in hodgkin and primary mediastinal B-cell lymphoma allows prediction of therapeutic targets,* Cancer Research **71**, 693 (2011).

[32] A. Raue, J. Karlsson, M. P. Saccomani, M. Jirstrand, and J. Timmer, *Comparison of approaches for parameter identifiability analysis of biological systems,* Bioinformatics **30**, 1440–1448 (2014).

[33] D. V. Raman, *On the Identifiability of Highly Parameterised Models of Physical Processes*, Ph.D. thesis, University of Oxford (2016).

[34] F. Bianconi, E. Baldelli, V. Ludovini, L. Crinó, A. Flacco, and P. Valigi, *Computational model of egfr and igf1r pathways in lung cancer: a systems biology approach for translational oncology,* Biotechnol Adv **Jan-Feb**, 142 (2012).

[35] F. Bianconi, E. Baldelli, V. Ludovini, L. Crinó, A. Flacco, and P. Valigi, *Egfr and Igf1r pathway in lung cancer,* (2012).

APPENDIX

Supplementary material: An efficient procedure to reparameterise structurally unidentifiable models ☆

D. Joubert[a,], J.D. Stigter[a], J. Molenaar[a]

[a] *Wageningen University & Research, Biometris, Mathematical and Statistical Methods Group, Wageningen, The Netherlands.*

1. Results

Example 1: Immunological model for mastitis in dairy cows

The reparameterised model contains 6 parameters instead of 7 (θ_1, θ_3, θ_5, ϕ_1, $x_1(0)$, ϕ_2):

$$\dot{x_1} = \theta_1 x_1 - x_1 \tilde{x}_2, \tag{1}$$

$$\dot{\tilde{x}}_2 = \theta_3 \tilde{x}_2 (1 - \phi_1 \tilde{x}_2) + \theta_5 x_1 \tilde{x}_2. \tag{2}$$

With $x_1(0)$ and $\tilde{x}_2(0) = \phi_2$.

This model is identifiable. This is shown in Figure 1, where we observe that none of the singular values are zero. In addition, Figure 2 shows that the reparameterised model and the original model predict the same output for state x_1.

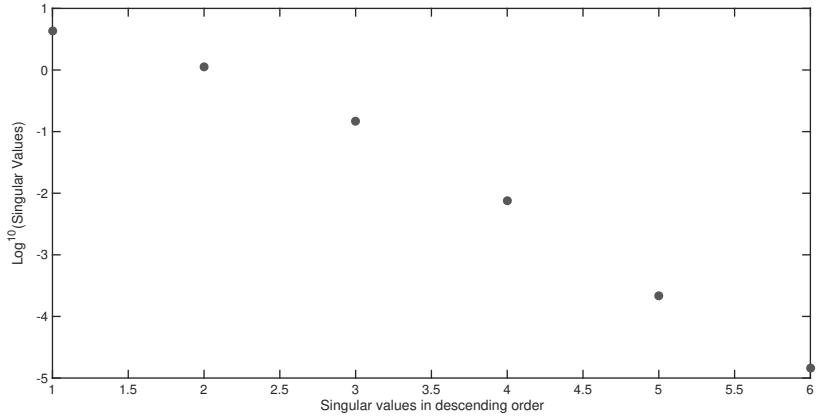


Figure 1: **Identifiability signature's singular values for the reparameterised Immunological model** - Singular values indicate that the model is identifiable since there is no zero-valued singular values and so the sensitivity matrix is of full rank. There are only 6 singular values since the total number of model parameters was reduced by 1, corresponding with the fact that there was only 1 set of correlated parameters in the original model.

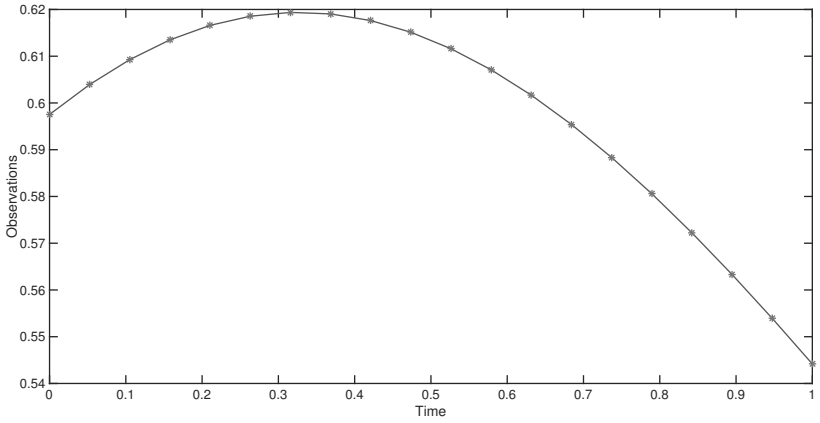


Figure 2: **Predicted model output of the original unidentifiable model and the reparameterised Immunological model** - solid line depicts the original model's output and $*$ depicts the new model's output. These graphs are perfectly aligned demonstrating that the 2 models predict similar outcomes.

Example 2: Microbial growth model

The reparameterised model contains 5 parameters instead of 6 (μ, K_d, ϕ_1, ϕ_2 and $x(0)$). The initial condition related to $s(0)$ now regarded as known, $\tilde{s}(0) = 1$.

$$\dot{x} = \frac{\mu x \tilde{s}}{\phi_1 + \tilde{s}} - K_d x, \tag{3}$$

$$\dot{\tilde{s}} = -\frac{\mu x \tilde{s}}{\phi_2(\phi_1 + \tilde{s})}. \tag{4}$$

This model is identifiable. This is shown in Figure 3, where we observe that none of the singular values are zero. In addition, Figure 4 shows that the reparameterised model and the original model predict the same output for state x.

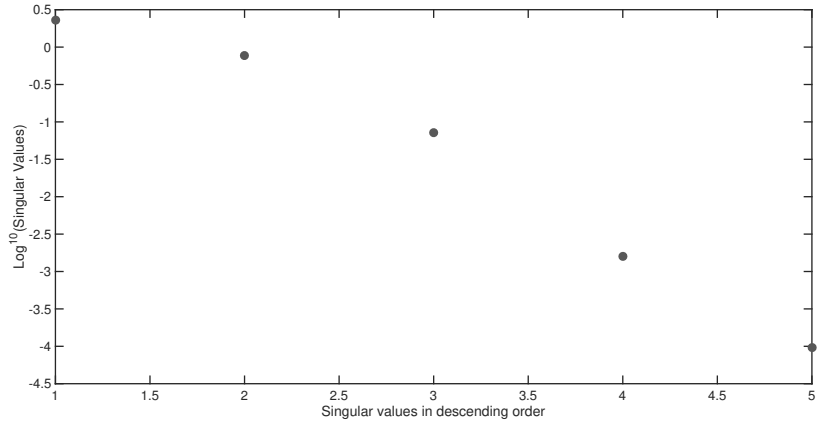


Figure 3: **Identifiability signature's singular values for the reparameterised Microbial growth model** - Singular values indicate that the model is identifiable since there is no zero-valued singular values and so the sensitivity matrix is of full rank.

The alternative model reparametrisation, were Y is removed from the model is also identifiable. The 5 unknown parameters of this reparameterisation are: $\mu, K_d, \phi_1, \phi_2 = s(0)$ and $x(0)$. The model:

$$\dot{x} = \frac{\mu x \tilde{s}}{\phi_2\phi_1 + \tilde{s}} - K_d x, \tag{5}$$

$$\dot{\tilde{s}} = -\frac{\mu x 1 \tilde{s}}{\phi_2\phi_1 + \tilde{s}}, \tag{6}$$

with initial conditions defied as $x(0)$ and $\tilde{s}(0) = \phi_2$ respectively. Newly computed singular values can be seen in figure 5.

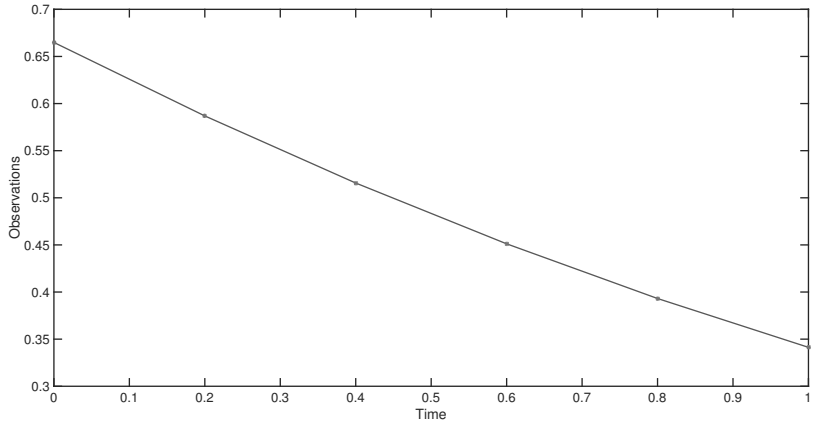


Figure 4: **Predicted model output of the original unidentifiable model and the reparameterised Microbial growth model** - solid line depicts the original model's output and $*$ depicts the new model's output. These graphs are perfectly aligned demonstrating that the 2 models predict similar outcomes.

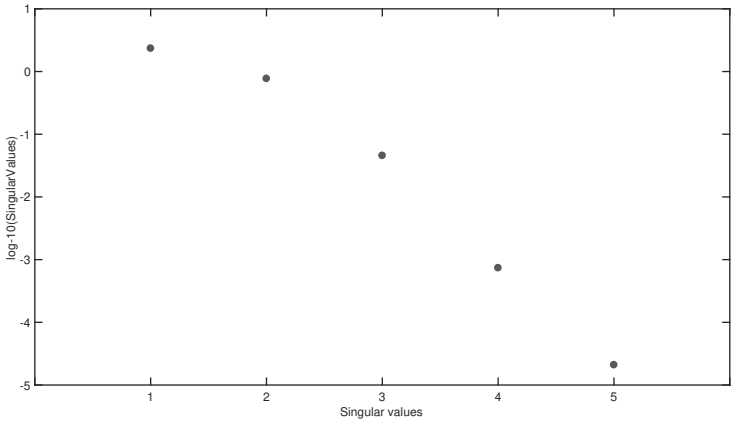


Figure 5: **Identifiability signature's singular values for the reparameterised Microbial growth model - Version 2** - Singular values indicate that the model is identifiable since there is no zero-valued singular values and so the sensitivity matrix is of full rank.

Example 3: JAK/STAT model

The reparameterised model contains 21 parameters instead of 23 ($\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8, \theta_9, \theta_{10}, \theta_{12}, \theta_{13}, \theta_{14}, \theta_{16}, \theta_{18}, \theta_{19}, \theta_{20}, \phi_{1,1}, \phi_{2,1}, \phi_{2,2}$ and $x_2(0) = \theta_{23}$) and reads as:

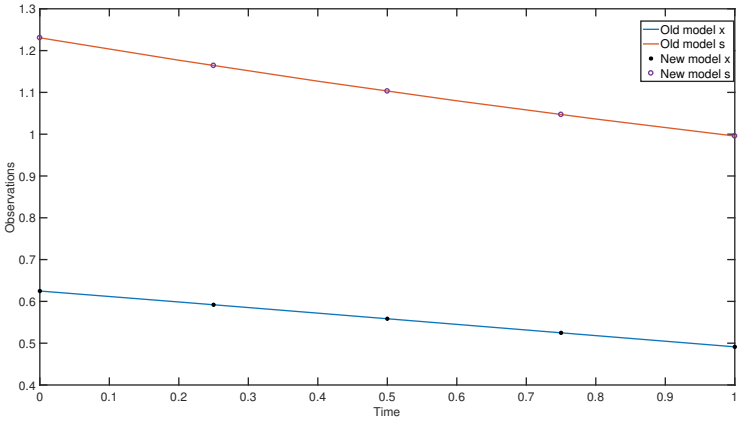


Figure 6: **Predicted model output of the original unidentifiable model and the reparameterised Microbial growth model - Version 2** - These graphs are perfectly aligned demonstrating that the 2 models predict similar outcomes.

$$\dot{x}_1 = -\theta_1 u_1 c_1 x_1 - \theta_5 x_1 + \theta_6 x_2, \tag{7}$$

$$\dot{x}_2 = \theta_5 x_1 - \theta_6 x_2, \tag{8}$$

$$\dot{x}_3 = \theta_1 u_1 c_1 x_1 - \theta_2 x_3 x_7, \tag{9}$$

$$\dot{x}_4 = \theta_2 x_3 x_7 - \theta_3 x_4, \tag{10}$$

$$\dot{x}_5 = \theta_3 x_4 - \theta_4 x_5, \tag{11}$$

$$\dot{x}_6 = -\frac{\theta_7 x_3 x_6}{(1+\theta_{13} x_{13})} - \frac{\theta_7 x_4 x_6}{(1+\theta_{13} x_{13})} + \theta_8 c_2 x_7, \tag{12}$$

$$\dot{x}_7 = \frac{\theta_7 x_3 x_6}{(1+\theta_{13} x_{13})} + \frac{\theta_7 x_4 x_6}{(1+\theta_{13} x_{13})} - \theta_8 c_2 x_7, \tag{13}$$

$$\dot{x}_8 = -\theta_9 x_8 x_7 + c_2 \theta_{10} x_9, \tag{14}$$

$$\dot{x}_9 = \theta_9 x_8 x_7 - c_2 \theta_{10} x_9, \tag{15}$$

$$\dot{\tilde{x}}_{10} = x_9, \tag{16}$$

$$\dot{x}_{11} = -\theta_{12} c_1 u_1 x_{11}, \tag{17}$$

$$\dot{x}_{12} = \theta_{12} c_1 u_1 x_{11}, \tag{18}$$

$$\dot{x}_{13} = \frac{\theta_{14} \tilde{x}_{10}}{(\phi_{2,1} + \tilde{x}_{10})} - \theta_{16} x_{13}, \tag{19}$$

$$\dot{\tilde{x}}_{14} = x_9. \tag{20}$$

The model output contains 5 additional parameters:

$$y_1 = x_1 + x_3 + x_4, \tag{21}$$
$$y_2 = \theta_{18}(x_3 + x_4 + x_5 + x_{12}), \tag{22}$$
$$y_3 = \theta_{19}(x_4 + x_5), \tag{23}$$
$$y_4 = \theta_{20}x_7, \tag{24}$$
$$y_5 = \phi_{2,2}\tilde{x}_{10}, \tag{25}$$
$$y_6 = \phi_{1,1}\tilde{x}_{14}, \tag{26}$$
$$y_7 = x_{13}, \tag{27}$$
$$y_8 = x_9. \tag{28}$$

In Figure 7 we show that this model has no gap between in singular values. Figure 8 shows that the reparameterised model and the original model yield identical outputs.

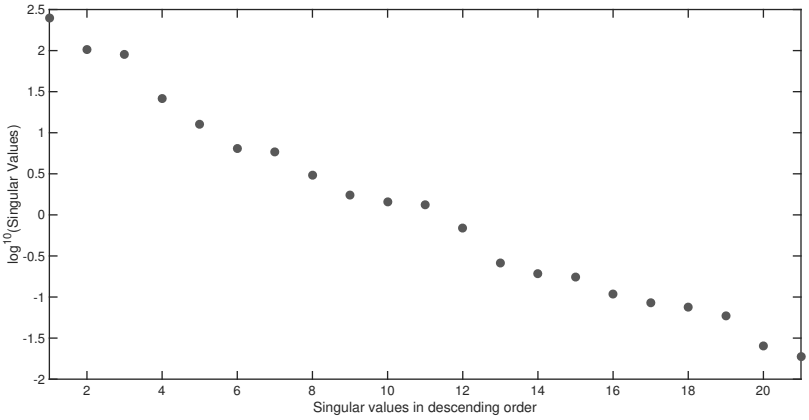


Figure 7: **Identifiability signature's singular values for the reparameterised JAK/STAT model** - Singular values indicate that the reparameterised model is identifiable since there is no zero-valued singular values and so the sensitivity matrix is of full rank.

Example 4: Lung cancer model

The original Lung cancer model:

d(EGFR_active)/dt = (-gamma_EGFR*EGFR_active)

d(D_SOS)/dt = (-k_SOS_E*EGFR_active*power(D_SOS,n_SOS)/(power(KM_SOS_E,n_SOS)+power(D_SOS,n_SOS)) + P90Rsk_Active*k_D_SOS_P90Rsk*power(A_SOS,n_D_SOS)/(power(KM_D_SOS_P90Rsk,n_D_SOS)+power(A_SOS,n_D_SOS)) - IGFR_active*k_A_SOS_I*power(D_SOS,n_A_SOS_I)/(power(KM_A_SOS_I,n_A_SOS_I)+power(D_SOS,n_A_SOS_I)))

d(A_SOS)/dt = (k_SOS_E*EGFR_active*power(D_SOS,n_SOS)/(power(KM_SOS_E,n_SOS)+power(D_SOS,n_SOS)) - P90Rsk_Active*k_D_SOS_P90Rsk*power(A_SOS,n_D_SOS)/(power(KM_D_SOS_P90Rsk,n_D_SOS)+power(A_SOS,n_D_SOS)) + IGFR_active*k_A_SOS_I*power(D_SOS,n_A_SOS_I)/(power(KM_A_SOS_I,n_A_SOS_I)+power(D_SOS,n_A_SOS_I)))

d(Raf)/dt = (-Ras_active*k_Raf_RasActive*power(Raf,n_Raf_RasActive)/(KM_Raf_RasActive+power(Raf,n_Raf_RasActive)) + AKT_active*k_Raf_AKT*power(Raf_active,n_Raf_AKT)/(power(KM_Raf_AKT,n_Raf_AKT)+power(Raf_active,n_Raf_AKT)) + RafPP*k_RasActive_RafPP*power(Raf_active,n_RasActive_RafPP)/(power(KM_RasActive_RafPP,n_RasActive_RafPP) +power(Raf_active,n_RasActive_RafPP)))

d(Ras_active)/dt = (A_SOS*k_Ras_SOS*power(Ras,n_Ras_SOS)/(power(KM_Ras_SOS,n_Ras_SOS)+power(Ras,n_Ras_SOS)) - RasGapActive*k_RasActiveRasGap*power(Ras_active,n_RasActiveRasGap)/(power(KM_RasActiveRasGap,n_RasActiveRasGap)+power(Ras_active,n_RasActiveRasGap)))

d(Mek_active)/dt =
(Raf_active*k_Mek_PP2A*power(Mek,n_Mek_PP2A)/(power(KM_MekPP2A,n_Mek_PP2A)+power(Mek,n_Mek_PP2A)) - PP2A*k_MekActivePP2A*power(Mek_active,n_MekActivePP2A)/(power(KM_MekActivePP2A,n_MekActivePP2A)+power(Mek_active,n_MekActivePP2A)))

d(ERK)/dt = (-Mek_active*k_ERK_MekActive*ERK/(KM_ERK_MekActive+ERK) + PP2A*k_ERKactive_PP2A*power(ERK_active,n_ERKactive_PP2A)/(power(KM_ERKactive_PP2A,n_ERKactive_PP2A)+power(ERK_active,n_ERKactive_PP2A)))

d(ERK_active)/dt = (Mek_active*k_ERK_MekActive*ERK/(KM_ERK_MekActive+ERK) - PP2A*k_ERKactive_PP2A*power(ERK_active,n_ERKactive_PP2A)/(power(KM_ERKactive_PP2A,n_ERKactive_PP2A)+power(ERK_active,n_ERKactive_PP2A)))

d(IGFR_active)/dt = (-gamma_IGFR*IGFR_active)

d(PI3KCA)/dt = (-
IGFR_active*k_PI3K_IGF1R*power(PI3KCA,n_PI3K_I)/(power(KM_PI3K_IGF1R,n_PI3K_I)+power(PI3KCA,n_PI3K_I)) - EGFR_active*k_PI3K_EGF1R*EGFR_active*power(PI3KCA,n_PI3K_E)/(power(KM_PI3K_EGF1R,n_PI3K_E)+power(PI3KCA,n_PI3K_E)) - Ras_active*k_PI3K_Ras*power(PI3KCA,n_PI3K_Ras)/(power(KM_PI3K_Ras,n_PI3K_Ras)+power(PI3KCA,n_PI3K_Ras)) + kd_PI3K_a*PI3KCA_active)

d(PI3KCA_active)/dt =
(IGFR_active*k_PI3K_IGF1R*power(PI3KCA,n_PI3K_I)/(power(KM_PI3K_IGF1R,n_PI3K_I)+power(PI3KCA,n_PI3K_I)) + EGFR_active*k_PI3K_EGF1R*EGFR_active*power(PI3KCA,n_PI3K_E)/(power(KM_PI3K_EGF1R,n_PI3K_E)+power(PI3KCA,n_PI3K_E)) +
Ras_active*k_PI3K_Ras*power(PI3KCA,n_PI3K_Ras)/(power(KM_PI3K_Ras,n_PI3K_Ras)+power(PI3KCA,n_PI3K_Ras)) - kd_PI3K_a*PI3KCA_active)

d(AKT_active)/dt =
(PI3KCA_active*k_AKT_PI3K*power(AKT,n_AKT_PI3K)/(power(KM_AKT_PI3K,n_AKT_PI3K)+power(AKT,n_AKT_PI3K)) - kd_AKT*AKT_active)

d(AKT)/dt = (-
PI3KCA_active*k_AKT_PI3K*power(AKT,n_AKT_PI3K)/(power(KM_AKT_PI3K,n_AKT_PI3K)+power(AKT,n_AKT_PI3K)) + kd_AKT*AKT_active)

d(PP2A)/dt =0

d(Ras)/dt = (-A_SOS*k_Ras_SOS*power(Ras,n_Ras_SOS)/(power(KM_Ras_SOS,n_Ras_SOS)+power(Ras,n_Ras_SOS)) + RasGapActive*k_RasActiveRasGap*power(Ras_active,n_RasActiveRasGap)/(power(KM_RasActiveRasGap,n_RasActiveRasGap)+power(Ras_active,n_RasActiveRasGap)))

d(Raf_active)/dt = (Ras_active*k_Raf_RasActive*power(Raf,n_Raf_RasActive)/(KM_Raf_RasActive+power(Raf,n_Raf_RasActive)) - AKT_active*k_Raf_AKT*power(Raf_active,n_Raf_AKT)/(power(KM_Raf_AKT,n_Raf_AKT)+power(Raf_active,n_Raf_AKT)) - RafPP*k_RasActive_RafPP*power(Raf_active,n_RasActive_RafPP)/(power(KM_RasActive_RafPP,n_RasActive_RafPP)+power(Raf_active,n_RasActive_RafPP)))

d(Mek)/dt = (-Raf_active*k_Mek_PP2A*power(Mek,n_Mek_PP2A)/(power(KM_MekPP2A,n_Mek_PP2A)+power(Mek,n_Mek_PP2A)) + PP2A*k_MekActivePP2A*power(Mek_active,n_MekActivePP2A)/(power(KM_MekActivePP2A,n_MekActivePP2A)+power(Mek_active,n_MekActivePP2A)))

d(RasGapActive)/dt=0

d(RafPP)/dt=0

d(P90RskInactive)/dt = (-ERK_active*k_P90Rsk_ERKActive*P90RskInactive/(KM_P90Rsk_ERKActive+P90RskInactive) + kd_P90Rsk*P90Rsk_Active)

d(P90Rsk_Active)/dt = (ERK_active*k_P90Rsk_ERKActive*P90RskInactive/(KM_P90Rsk_ERKActive+P90RskInactive) - kd_P90Rsk*P90Rsk_Active)

We re-write this model in terms of x (for model states) and th (for system parameters):

State name	Symbol
EGFR_active	x1
D_SOS	x2
A_SOS	x3
Raf	x4
Ras_active	x5
Mek_active	x6
ERK	x7
ERK_active	x8
IGFR_active	x9
PI3KCA	x10
PI3KCA_active	x11
AKT_active	x12
AKT	x13
PP2A	x14
Ras	x15
Raf_active	x16
Mek	x17
RasGapActive	x18
RafPP	x19
P90RskInactive	x20
P90Rsk_Active	x21

Parameter name	Symbol
gamma_IGFR	th1
kd_PI3K_a	th2
k_P90Rsk_ERKActive	th3
KM_P90Rsk_ERKActive	th4
gamma_EGFR	th5
k_SOS_E	th6
n_SOS	th7
KM_SOS_E	th8
k_Ras_SOS	th9
n_Ras_SOS	th10
KM_Ras_SOS	th11
k_ERK_MekActive	th12
KM_ERK_MekActive	th13
k_D_SOS_P90Rsk	th14
n_D_SOS	th15
KM_D_SOS_P90Rsk	th16
k_A_SOS_I	th17
n_A_SOS_I	th18
KM_A_SOS_I	th19
k_PI3K_IGF1R	th20
n_PI3K_I	th21
KM_PI3K_IGF1R	th22
k_PI3K_EGF1R	th23
n_PI3K_E	th24
KM_PI3K_EGF1R	th25
k_AKT_PI3K	th26
n_AKT_PI3K	th27
KM_AKT_PI3K	th28
kd_AKT	th29
k_ERKactive_PP2A	th30
n_ERKactive_PP2A	th31
KM_ERKactive_PP2A	th32
k_PI3K_Ras	th33
n_PI3K_Ras	th34
KM_PI3K_Ras	th35
k_Raf_RasActive	th36
n_Raf_RasActive	th37
KM_Raf_RasActive	th38
k_Mek_PP2A	th39
n_Mek_PP2A	th40
KM_MekPP2A	th41
k_Raf_AKT	th42
n_Raf_AKT	th43
KM_Raf_AKT	th44
k_RasActiveRasGap	th45
n_RasActiveRasGap	th46
KM_RasActiveRasGap	th47
k_MekActivePP2A	th48
n_MekActivePP2A	th49
KM_MekActivePP2A	th50

k_RasActive_RafPP	th51
n_RasActive_RafPP	th52
KM_RasActive_RafPP	th53
kd_P90Rsk	th54

Old unidentifiable model:

d(x1)/dt = -th5*x1

d(x2)/dt = -th6*x1*x2^th7/(th8^th7+x2^th7) + x21*th14*x3^th15/(th16^th15+x3^th15) - x9*th17*x2^th18/(th19^th18+x2^th18)

d(x3)/dt = th6*x1*x2^th7/(th8^th7+x2^th7) - x21*th14*x3^th15/(th16^th15+x3^th15) + x9*th17*x2^th18/(th19^th18+x2^th18)

d(x4)/dt = -x5*th36*x4^th37/(th38+x4^th37) + x12*th42* x16^th43/(th44^th43+ x16^th43) + x19*th51* x16^th52/(th53^th52+ x16^th52)

d(x5)/dt = x3*th9* x15^th10/(th11^th10+ x15^th10) – x18*th45*x5^th46/(th47^th46+x5^th46)

d(x6)/dt = (x16*th39* x17^th40/(th41^th40+ x17^th40) - x14*th48*x6^th49/(th50^th49+ x6^th49))

d(x7)/dt = x14*th30* x8^th31/(th32^th31+x8^th31)-(x6*th12*x7)/(th13+x7)

d(x8)/dt = (x6*th12*x7)/(th13+x7)- x14*th30*x8^th31/(th32^th31+ x8^th31)

d(x9)/dt = -th1*x9

d(x10)/dt = (-x9*th20* x10^th21/(th22^th21+ x10^th21) - x1*th23*x1*x10^th24/(th25^th24+ x10^th24) - x5*th33* x10^th34/(th35^th34+ x10^th34) + th2*x11)

d(x11)/dt = (x9*th20* x10^th21/(th22^th21+ x10^th21) + x1*th23*x1*x10^th24/(th25^th24+ x10^th24) + x5*th33* x10^th34/(th35^th34+ x10^th34) – th2*x11)

d(x12)/dt = (x11*th26*x13^th27/(th28^th27+x13^th27) – th29*x12)

d(x13)/dt = (-x11*th26*x13^th27/(th28^th27+ x13^th27) + th29*x12)

d(x14)/dt=0

d(x15)/dt = (-x3*th9* x15^th10/(th11^th10+ x15^th10) + x18*th45*x5^th46/(th47^th46+x5^th46))

d(x16)/dt = (x5*th36* x4^th37/(th38+x4^th37) - x12*th42* x16^th43/(th44^th43+ x16^th43) - x19*th51* x16^th52/(th53^th52+ x16^th52))

d(x17)/dt = (-x16*th39* x17^th40/(th41^th40+ x17^th40) + x14*th48* x6^th49/(th50^th49+ x6^th49))

d(x18)/dt=0

d(x19)/dt=0

d(x20)/dt = (-x8*th3* x20/(th4+ x20) + th54*x21)

d(x21)/dt = (x8*th3* x20/(th4+ x20) – th54*x21)

The reparameterized identifiable model contains 73 parameters:

(th1, th2, th3, th4, th5, th7, th8, th9, th10, th11, th12, th13, th14, th15, th16, th18, th19, th21, th22, th24, th25, th26, th27, th28, th29, th30, th31, th32, th33, th34, th35, th36, th37, th38, th39, th40, th41, th42, th43, th44, th45, th46, th47, th48, th49, th50, th51, th52, th53, th54, $\emptyset_{1,1}$, $\emptyset_{2,1}$, and initial conditions $\emptyset_{2,2}$, x2(0), x3(0), x4(0), x5(0), x6(0), x7(0), x8(0), $\emptyset_{1,2}$, x10(0), x11(0), x12(0), x13(0), x14(0), x15(0), x16(0), x17(0), x18(0), x19(0), x20(0), x21(0))

d($\tilde{x}1$)/dt = -th5*$\tilde{x}1$

d(x2)/dt = -$\tilde{x}1$*x2^th7/(th8^th7+x2^th7) + x21*th14*x3^th15/(th16^th15+x3^th15) - $\tilde{x}9$*x2^th18/(th19^th18+x2^th18)

d(x3)/dt = $\tilde{x}1$*x2^th7/(th8^th7+x2^th7) - x21*th14*x3^th15/(th16^t``h15+x3^th15) + $\tilde{x}9$*x2^th18/(th19^th18+x2^th18)

d(x4)/dt = -x5*th36*x4^th37/(th38+x4^th37) + x12*th42* x16^th43/(th44^th43+ x16^th43) + x19*th51* x16^th52/(th53^th52+ x16^th52)

d(x5)/dt = x3*th9* x15^th10/(th11^th10+ x15^th10) – x18*th45*x5^th46/(th47^th46+x5^th46)

d(x6)/dt = (x16*th39* x17^th40/(th41^th40+ x17^th40) - x14*th48*x6^th49/(th50^th49+ x6^th49))

d(x7)/dt = x14*th30* x8^th31/(th32^th31+x8^th31)-(x6*th12*x7)/(th13+x7)

d(x8)/dt = (x6*th12*x7)/(th13+x7)- x14*th30*x8^th31/(th32^th31+ x8^th31)

d($\tilde{x}9$)/dt = -th1$\tilde{x}9$

d(x10)/dt = (-$\tilde{x}9$*$\emptyset_{1,1}$* x10^th21/(th22^th21+ x10^th21) - $\tilde{x}1$*$\emptyset_{2,1}$*$\tilde{x}1$*x10^th24/(th25^th24+ x10^th24) - x5*th33* x10^th34/(th35^th34+ x10^th34) + th2*x11)

d(x11)/dt = ($\tilde{x}9$*$\emptyset_{1,1}$* x10^th21/(th22^th21+ x10^th21) + $\tilde{x}1$*$\emptyset_{2,1}$*$\tilde{x}1$*x10^th24/(th25^th24+ x10^th24) + x5*th33* x10^th34/(th35^th34+ x10^th34) – th2*x11)

d(x12)/dt = (x11*th26*x13^th27/(th28^th27+x13^th27) – th29*x12)

d(x13)/dt = (-x11*th26*x13^th27/(th28^th27+ x13^th27) + th29*x12)

d(x14)/dt=0

d(x15)/dt = (-x3*th9* x15^th10/(th11^th10+ x15^th10) + x18*th45*x5^th46/(th47^th46+x5^th46))

d(x16)/dt = (x5*th36* x4^th37/(th38+x4^th37) - x12*th42* x16^th43/(th44^th43+ x16^th43) - x19*th51* x16^th52/(th53^th52+ x16^th52))

d(x17)/dt = (-x16*th39* x17^th40/(th41^th40+ x17^th40) + x14*th48* x6^th49/(th50^th49+ x6^th49))

d(x18)/dt=0

d(x19)/dt=0

d(x20)/dt = (-x8*th3* x20/(th4+ x20) + th54*x21)

d(x21)/dt = (x8*th3* x20/(th4+ x20) – th54*x21)

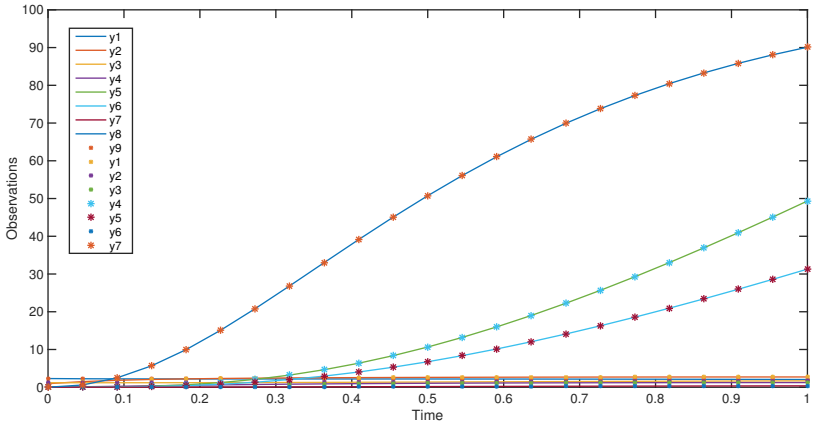


Figure 8: **Outputs of the original unidentifiable model and the reparameterised JAK/STAT model** - solid lines depict the original model's outputs and $*$ depict the new model's outputs. These graphs are perfectly aligned demonstrating that the 2 models predict similar outcomes.

This model is identifiable. This is shown in Figure 9, where we observe that none of the singular values are zero. In addition, Figure 10 and 11 show that the reparameterised model and the original model predict the same output.

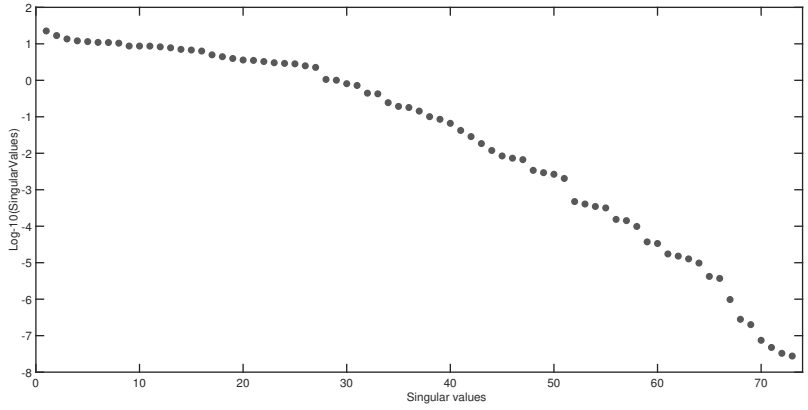


Figure 9: **Identifiability signature's singular values for the reparameterised Lung cancer model** - Singular values indicate that the reparameterised model is identifiable since there is no zero-valued singular values and so the sensitivity matrix is of full rank.

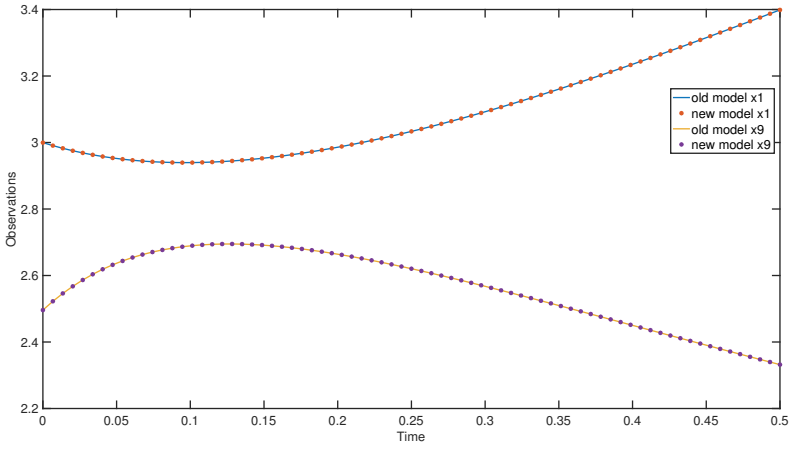


Figure 10: **Outputs x_1 and x_9 of the original unidentifiable model and the reparameterised Lung cancer model** - solid lines depict the original model's outputs and $*$ depict the new model's outputs. These graphs are perfectly aligned demonstrating that the 2 models predict similar outcomes.

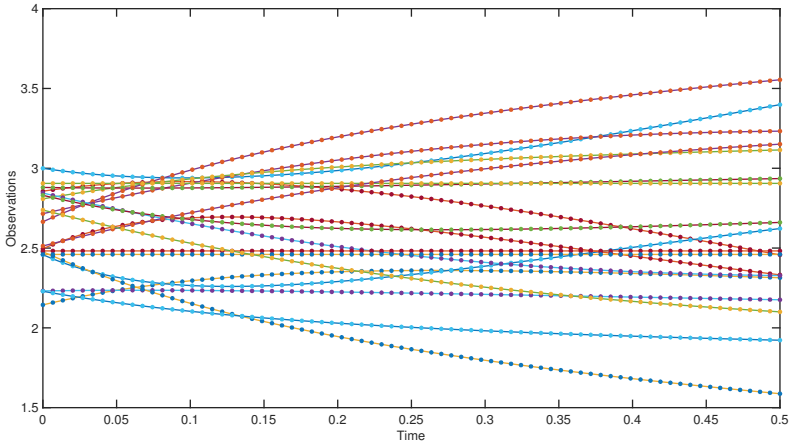


Figure 11: **Outputs of the original unidentifiable model and the reparameterised Lung cancer model** - solid lines depict the original model's outputs and $*$ depict the new model's outputs. These graphs are perfectly aligned demonstrating that the 2 models predict similar outcomes.

Directed graphs of original models
JAK/STAT model

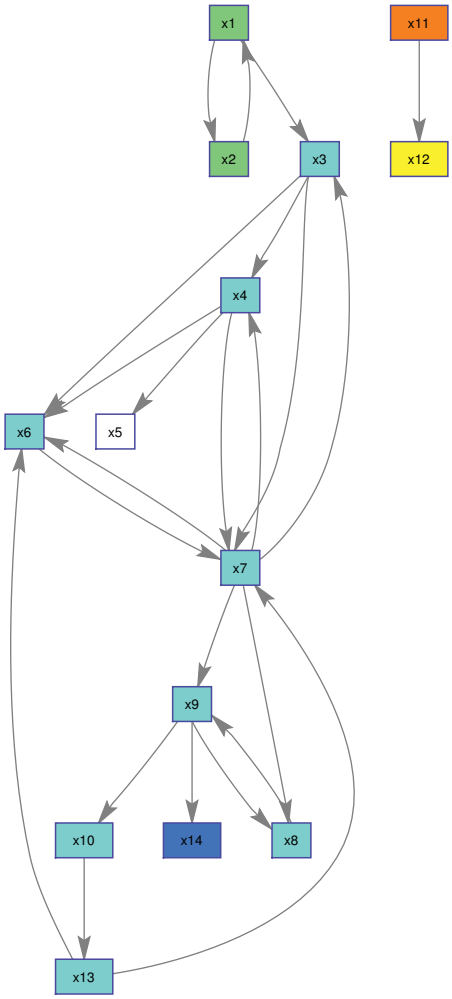


Figure 12: **Directed graph of the JAK/STAT model**

Lung cancer model

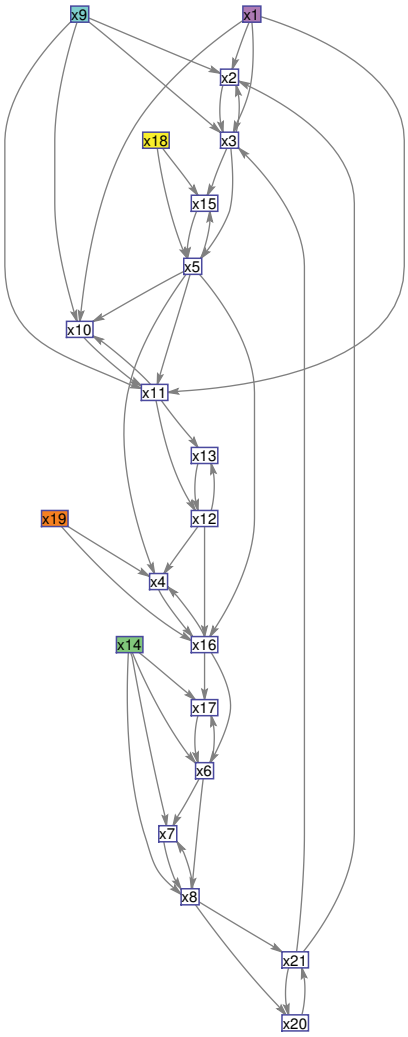


Figure 13: **Directed graph of the Lung cancer model**

4

NUMERICAL SENSITIVITY OF THE LOCAL STRUCTURAL IDENTIFIABILITY ALGORITHM

Dominique JOUBERT, Hans STIGTER, Jaap MOLENAAR

Despite the interest in knowing a priori whether there is any chance of uniquely estimating all model unknown parameters, the structural identifiability analysis for general nonlinear dynamic models is still an open question.
There is no method amenable to every model...

(Chis, Banga, and Balsa-Canto, 2011)

ABSTRACT

THE increase in the number of large and complex ODE models in the field of systems biology continues to drive the development of fast and user friendly methods that are capable of analysing these models. In chapters 2 and 3 we saw that our hybrid SVD approach is efficient in terms of computation time and therefore lends itself to additional model analyses such as, determining minimal output sets. In this chapter, we pin down some standard settings that can be used in future software implementations of the algorithm. Since there exists no single method that is amenable to every model, the aim is to have a robust method at one's disposal, capable of analysing a wide range of models. Here, we show that our algorithm is capable of this.

In this chapter, the sensitivity of the identifiability results with respect to user adjustable settings is methodically investigated. Results show that *parameter and initial values*, in particular, are influential. In addition, a model's size dictates whether symbolic or numerical matrices, $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}$ and $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\theta}}$, can be computed and this may influence the sharpness results for some models. A third important factor is the *length* of the output vector. Finally, we show that the vertical concatenation of different sensitivity matrices considerably reduces the influence of both parameter values and output vector length. The robustness of our method is investigated analysing seven examples, and we find that for the majority of models, the implementation of standard settings yield accurate results.

Highlights include: 1) the introduction of rule of thumb settings that can be implemented when first analysing a model and, 2) illustrating the advantage of vertically concatenating different sensitivity matrices in obtaining decisive unidentifiability results.

4.1. INTRODUCTION

The drive to incorporate vast amounts of knowledge into a single model will in the future be reflected in large numbers of complex mathematical systems. In the field of systems biology, a model may for example predict the movement of a certain cancer drug into a target cell and therefore focus on a very specific process. In such cases, models can be small. Alternatively, a set of smaller models can be incorporated into a larger, more holistic model. These larger models may be tasked with describing a cell in its entirety. Given the potential applications of systems biology models in the medical field, model accuracy may indeed be the defining factor between life or death. Consequently, there is a need for tools that can analyse the accuracy of large models in particular.

Structural identifiability is the starting point when analysing a model's predictive power. It determines, *a priori*, whether the model's structure will allow for the calculation of unique parameter values. This property is traditionally analysed symbolically, but with the surge of larger models, numerical identifiability methods seem to be a logical step towards addressing this problem. Yet, questions regarding these methods do exist and in this chapter we aim to address some of these questions.

The motivation for writing this chapter is summarised in the following quote:

"*Numerical analysis is heavily dependent on notional values for the parameter (that are to be estimated), and involves applying sampling rate to the output. These results are therefore affected by a number of factors that one would wish to understand the individual effect of – for example, is a model over-parameterised regardless of the number and timing of samples taken*" (Evans, Cheung, Yates, 2018) [1].

In the following sections, we will analyse the different factors that influence numerical identifiability results. This analysis can be framed as an investigation into the robustness of the method used in this thesis. The aim of this chapter is to understand the influence these factors have on two important properties: 1) the size of the gap in singular values. The larger this gap, the more decisive the unidentifiability result, and 2) numerical integration times, with shorter times allowing for more diverse applications of the method. The obtained outcomes in the future play an important role in the development of a software application, and will ensure that users obtain reliable numerical identifiability results. The results of this work will be a number of guidelines for analysis settings, in other words, a set of rules of thumb.

4.2. METHOD DESCRIPTION

The models analysed here are nonlinear autonomous ODE systems of the form:

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x}(t), \boldsymbol{\theta}), \tag{4.1}$$

$$\boldsymbol{x}(0) = \boldsymbol{x_0}, \tag{4.2}$$

$$\boldsymbol{y}(t) = \boldsymbol{h}(\boldsymbol{x}(t), \boldsymbol{\theta}). \tag{4.3}$$

Function $\boldsymbol{f}$ denotes a dynamic model structure and $\boldsymbol{h}$ the output or observation function. State variables are contained in vector $\boldsymbol{x}(t)$ ($\dim(\boldsymbol{x}) = n$), parameters in vector $\boldsymbol{\theta}$ ($\dim(\boldsymbol{\theta}) = p$) and model outputs, also described as measured sensors, in vector $\boldsymbol{y}(t)$

($\dim(\boldsymbol{y}) = m$). Initial values of the model states can also be unknown and in such instances we regard them as additional unknown parameters. In which case a parameter vector has $\dim(\boldsymbol{\theta}) = p + n$.

Sensitivity functions are obtained calculating the sensitivity of elements in the output vector, $\boldsymbol{y}$, with respect to individual parameters in $\boldsymbol{\theta}$. These sensitivities are calculated by differentiating equations 4.1 and 4.3 with respect to $\boldsymbol{x}$ and $\boldsymbol{\theta}$ and, numerically integrating the newly defined equation 4.4 with respect to time:

$$\frac{d}{dt}\left(\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\theta}}\right) = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\theta}} + \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\theta}}, \tag{4.4}$$

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{\theta}} = \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}}\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\theta}} + \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{\theta}}. \tag{4.5}$$

4

Individual sensitivities, calculated at discretized time points on a finite time interval $[0, t_N]$, are vertically concatenated to form the so-called sensitivity matrix, $\boldsymbol{S}$. If the initial conditions of model states are also unknown, $\boldsymbol{S}$ has $p+n$ columns, each related to an unknown parameter.

$$\boldsymbol{S} = \begin{pmatrix} \frac{\partial y_1}{\partial \theta_1}(t_0) & \cdots & \frac{\partial y_1}{\partial \theta_{p+n}}(t_0) \\ \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial \theta_1}(t_0) & \cdots & \frac{\partial y_m}{\partial \theta_{p+n}}(t_0) \\ \vdots & & \vdots \\ \frac{\partial y_1}{\partial \theta_1}(t_N) & \cdots & \frac{\partial y_1}{\partial \theta_{p+n}}(t_N) \\ \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial \theta_1}(t_N) & \cdots & \frac{\partial y_m}{\partial \theta_{p+n}}(t_N) \end{pmatrix}. \tag{4.6}$$

Rows and columns in this matrix can be normalised without changing its rank. This is typically done to reduce the effect of scaling differences on the numerical results. The normalised matrix used in this thesis is:

$$\boldsymbol{S}_{norm} = \begin{pmatrix} \frac{\theta_1}{y_1(t_0)}\frac{\partial y_1}{\partial \theta_1}(t_0) & \cdots & \frac{\theta_{p+n}}{y_1(t_0)}\frac{\partial y_1}{\partial \theta_{p+n}}(t_0) \\ \vdots & \ddots & \vdots \\ \frac{\theta_1}{y_m(t_0)}\frac{\partial y_m}{\partial \theta_1}(t_0) & \cdots & \frac{\theta_{p+n}}{y_m(t_0)}\frac{\partial y_m}{\partial \theta_{p+n}}(t_0) \\ \vdots & & \vdots \\ \frac{\theta_1}{y_1(t_N)}\frac{\partial y_1}{\partial \theta_1}(t_N) & \cdots & \frac{\theta_{p+n}}{y_1(t_N)}\frac{\partial y_1}{\partial \theta_{p+n}}(t_N) \\ \vdots & \ddots & \vdots \\ \frac{\theta_1}{y_m(t_N)}\frac{\partial y_m}{\partial \theta_1}(t_N) & \cdots & \frac{\theta_{p+n}}{y_m(t_N)}\frac{\partial y_m}{\partial \theta_{p+n}}(t_N) \end{pmatrix}. \tag{4.7}$$

Sensitivities are calculated for a particular set of parameter and initial values. To ensure that the model is analysed in multiple parameter trajectories, analyses are done for different sets of $\boldsymbol{\theta}^i$, where $i = 1,..,k$. This results in k different sensitivity matrices, $\boldsymbol{S}^i$ or $\boldsymbol{S}^i_{norm}$. These matrices can be vertically concatenated, resulting in a sensitivity matrix of the form:

$$\boldsymbol{S}_{all} = \begin{pmatrix} \boldsymbol{S}^1 \\ \boldsymbol{S}^2 \\ \vdots \\ \boldsymbol{S}^k \end{pmatrix}, \tag{4.8}$$

or

$$\boldsymbol{S}_{all,norm} = \begin{pmatrix} \boldsymbol{S}^1_{norm} \\ \boldsymbol{S}^2_{norm} \\ \vdots \\ \boldsymbol{S}^k_{norm} \end{pmatrix}. \tag{4.9}$$

The process of vertical concatenation is discussed in section 4.7. In our approach, an SVD is applied to any of the matrices described above (4.6, 4.7, 4.8 or 4.9). For example, if $\boldsymbol{S}$ is used, this matrix is decomposed as $\boldsymbol{S} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}$, and we use matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{V}$ in our analysis. $\boldsymbol{\Sigma}$ contains singular values on its diagonal and these are used to determine the rank of the sensitivity matrix, with rank-deficiency indicated by the presence of zero-valued singular values [2]. Due to numerical rounding errors, singular values are rarely exactly zero and accordingly, we use the following practical definition: Zero-valued singular values fall beyond a distinct gap in the spectrum of singular values [3]. In this thesis, we regard a gap larger than 3 decades as significant. A model's unidentifiability is indicated by a rank deficient sensitivity matrix and once possible unidentifiability has been established, unidentifiable parameters are recognised as the nonzero entries in the columns of the matrix $\boldsymbol{V}$, related to these vanishing singular values.

All numerical results are verified symbolically. However, the computational demand of these symbolic computations is significantly reduced by the fact that the Jacobi matrix only has to be computed for the unidentifiable parameters suggested by the numerical analysis. In the symbolic approach we make use of the Jacobi matrix in equation 4.12, with a nontrivial null-space of the Jacobi matrix indicating model unidentifiability. It contains the partial derivatives of the separate terms in the generating series of $\boldsymbol{h}$ with respect to $\boldsymbol{\theta}$. The individual terms of this series are calculated by computing successive Lie derivatives of the vector function $\boldsymbol{h}$ along the vector field, $\boldsymbol{f}$ [4]. The Lie derivative and its higher order derivatives are defined in equations 4.10 and 4.11 respectively [5].

$$\mathscr{L}_{\boldsymbol{f}}\boldsymbol{h}(\boldsymbol{x}) = \frac{\partial \boldsymbol{h}(\boldsymbol{x})}{\partial \boldsymbol{x}} \boldsymbol{f}(\boldsymbol{x}), \tag{4.10}$$

$$\mathscr{L}^i_{\boldsymbol{f}}\boldsymbol{h}(\boldsymbol{x}) = \frac{\partial \mathscr{L}^{i-1}_{\boldsymbol{f}}\boldsymbol{h}(\boldsymbol{x})}{\partial \boldsymbol{x}} \boldsymbol{f}(\boldsymbol{x}). \tag{4.11}$$

$$\frac{\partial \boldsymbol{G}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}) = \begin{pmatrix} \dfrac{\partial \boldsymbol{h}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\ \dfrac{\partial}{\partial \boldsymbol{\theta}}(\mathscr{L}_f \boldsymbol{h}(\boldsymbol{\theta})) \\ \dfrac{\partial}{\partial \boldsymbol{\theta}}(\mathscr{L}_f^2 \boldsymbol{h}(\boldsymbol{\theta})) \\ \vdots \\ \dfrac{\partial}{\partial \boldsymbol{\theta}}(\mathscr{L}_f^{p+n-1} \boldsymbol{h}(\boldsymbol{\theta})) \end{pmatrix}. \tag{4.12}$$

In this chapter we answer a number of questions regarding our numerical approach to the identifiablity problem:

4

1. Which integration interval, $[0, t_N]$, should be used?
2. Which numerical integration method is sufficient?
3. Which sampling frequency, translated as the time grid discretization, N, should be used when integrating? N stipulates the number of points in the interval $[0, t_N]$.
4. Which of the sensitivity matrices defined in 4.6-4.9, should be analysed?
5. Which parameter values and initial conditions should be used?
6. What potential effect does the length of the output vector, $\boldsymbol{y}$, and its contents have on results? The length of this vector is determined by the number of states/sensors measured.

This chapter is divided into the following sections. Potential problems that might be encountered when using the algorithm are mentioned in section 4.3. Factors that a user can adjust are listed in section 4.4. Detailed results are given in section 4.5 and the topic of matrix concatenation is covered in section 4.7. Concluding remarks can be found in the last section.

4.3. POTENTIAL PROBLEMS

In this section, we present some pitfalls that one might encounter when analysing a model. In addition, we suggest some practical tips for avoiding these instances.

4.3.1. SCALING

Pitfall: Results may be affected by the choice of parameter and initial values. To illustrate this, consider the identifiability analysis of a small ODE model with 1 state equation and 2 unknown system parameters. The model structure comprises the well-known Michealis-Menten equation [6, 7]:

$$\dot{x} = \frac{V_{max} x}{K_M + x}. \tag{4.13}$$

Unknown parameters, $\boldsymbol{\theta} = [V_{max}, K_M, x(0)]$, need to be estimated from the measured dynamic output, $y = x$. We start by randomly generating values for V_{max}, K_M and $x(0)$ in the

interval [0.5, 1.5]. With this choice, numerical results reveal that the model is structurally identifiable, with no significant gap between the singular values (figure 4.1). These results are confirmed by a symbolically calculated trivial null-space.

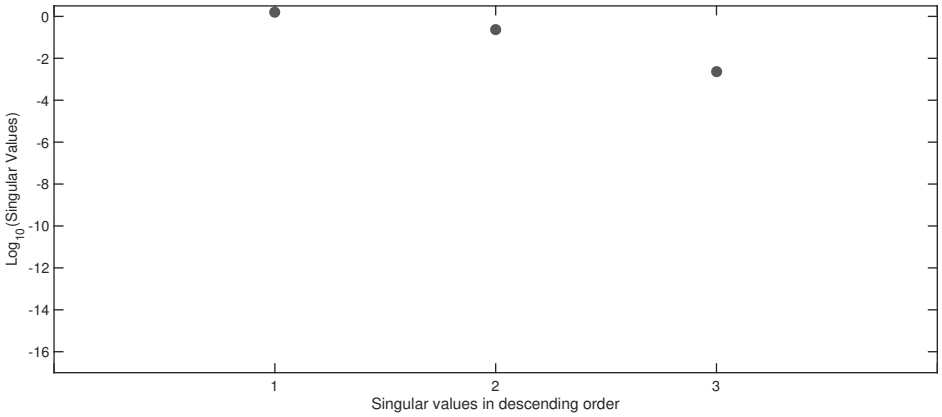


Figure 4.1: **Calculated singular values for the Michealis-Menten model defined in 4.13:** When parameter values are chosen in the interval [0.5, 1.5], no significant gap between the singular values is detected and the model is classified as structurally identifiable.

However, when the value of K_M is chosen much larger than that of V_{max} or $x(0)$, the obtained results differ significantly. When $V_{max} = 1, x(0) = 1$ and $K_M = 10^4$ for example, the largest gap size is 10.7 decades on the log scale, as seen in figure 4.2. There are 2 singular values beyond a gap larger than 3 and these urge us to look at the nonzero entries of the last 2 columns of the right singular matrix, $\boldsymbol{V}$, which are related to these 2 vanishing singular values. These 2 columns are shown in 4.14, and the first and second entries in each column suggest that both V_{max} and K_M are unidentifiable.

$$\boldsymbol{V} = \begin{pmatrix} -0.7071 & 0.7071 \\ 0.7071 & 0.7071 \\ 0 & 0 \end{pmatrix}. \tag{4.14}$$

To establish the identifiability of the model symbolically, the Jacobi matrix in equation 4.15 is calculated for all 3 parameters and up to the second Lie derivative using Mathematica.

$$\frac{d\boldsymbol{G}}{d\boldsymbol{\theta}}(\boldsymbol{\theta}) = \begin{pmatrix} 0 & 0 & 1 \\ \frac{x(0)}{(K_M+x(0))} & -\frac{V_{max}x(0)}{(K_M+x(0))^2} & -\frac{V_{max}x(0)}{(K_M+x(0))^2} + \frac{V_{max}}{K_M+x(0)} \\ \frac{2V_{max}K_Mx(0)}{(K_M+x(0))^3} & -\frac{3V_{max}^2K_Mx(0)}{(K_M+x(0))^4} + \frac{V_{max}^2x(0)}{(K_M+x(0))^3} & -\frac{3V_{max}^2K_Mx(0)}{(K_M+x(0))^4} + \frac{V_{max}^2K_M}{(K_M+x(0))^3}. \end{pmatrix}. \tag{4.15}$$

The determinant of this matrix is:

$$Det(\frac{d\boldsymbol{G}}{d\boldsymbol{\theta}}(\boldsymbol{\theta})) = -\frac{4V_{max}^4K_M^2x(0)^4}{(K_M+x(0))^{10}}, \tag{4.16}$$

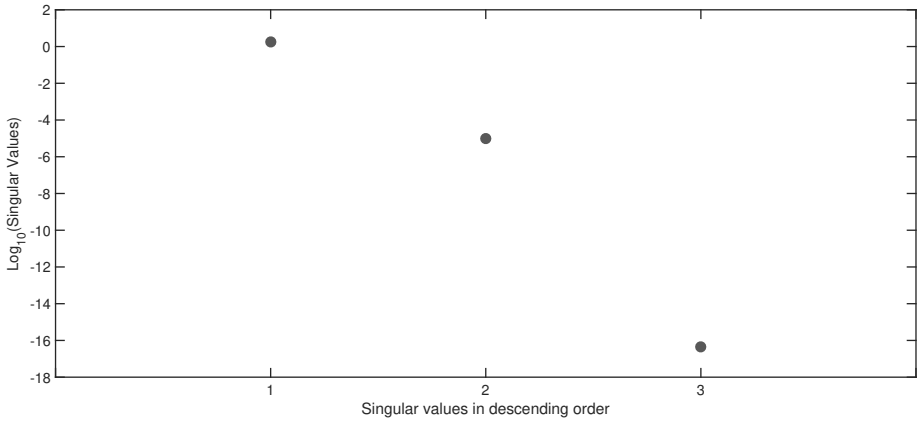


Figure 4.2: **Calculated singular values for the Michealis-Menten model defined in 4.13 when the parameter value for K_M is much larger than that of V_{max} and $x(0)$:** For parameter values $V_{max} = 1$, $K_M = 10^4$ and $x(0) = 1$, there are 2 singular values beyond a gap larger than 3 decades and therefore the model is incorrectly classified as structurally unidentifiable.

and substituting positive parameter values into equation 4.16, results in a nonzero determinant. This proves that the model is structurally identifiable as long as the parameter values are nonzero.

Practical tips:

1. Cautiously choose parameter values. Although unidentifiability is a model property that, except for special cases, does not depend on the parameter values used, numerical inaccuracy may blur this property. We recommend using values in the same order of magnitude. For example, randomly generated values between 0.5 and 1.5.

2. Pay special attention when models contain particularly large or small constants.

3. Verify all numerical results symbolically. In general, this does not require a great deal of computation, since we only need to analyse the identifiability of a reduced set of parameters. For this example, one would only need to verify the identifiability of parameters V_{max} and K_M symbolically.

4. The sensitivity matrix could be normalised. However, for this example this appears to be insufficient and accordingly, results should always be verified symbolically. If some of the initial conditions are zero, normalisation is even not possible (as seen in equation 4.7).

5. Repeat the analysis for multiple parameter trajectories.

4.3.2. STIFF ODE SYSTEMS

Pitfall: Stiff ordinary differential equations are often encountered in models that contain chemical reactions. In general, a problem is stiff if it contains widely varying time scales where, for example, some states decay much more rapidly than others. In these cases, numerical integration methods must take small steps to obtain satisfactory results and this increases both computational cost and time. In addition, longer integration times are required to ensure that all the model dynamics is captured.

As an example, consider the stiff Robertson ODE model with 3 state equations [8]:

$$\frac{dx_1}{dt} = -\theta_1 x_1 + \theta_2 x_2 x_3, \tag{4.17}$$

$$\frac{dx_2}{dt} = \theta_1 x_1 - \theta_2 x_2 x_3 - \theta_3 x_2^2, \tag{4.18}$$

$$\frac{dx_3}{dt} = \theta_3 x_2^2. \tag{4.19}$$

Parameter values are taken as $\theta_1 = 0.04, \theta_2 = 10^4$ and $\theta_3 = 3 \times 10^7$ respectively. Randomly generating values between 0.5 and 1.5 for the 3 initial conditions, and substituting both sets of values into the symbolically calculated Jacobi matrix, $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}$, the calculated real parts of the eigenvalues of this Jacobi matrix are:

$$\begin{pmatrix} -9.1524 \times 10^7 \\ -1.5254 \times 10^4 \\ 1.1362 \times 10^{-10} \end{pmatrix}. \tag{4.20}$$

The span in these eigenvalues indicate that this system is stiff and strongly suggests that one component evolves much slower than the others. Accordingly, one would require a long integration interval, roughly estimated as $t_N = \frac{1}{|\lambda_{Max}|} = \frac{1}{1.1362 \times 10^{-10}} = 8.801 \times 10^9$, to capture all the relevant dynamics. A specialised numerical integrator, for example ODE15s or ODE23s in MATLAB, determines the optimal size and number of subintervals into which the integration interval should be divided. These methods allow for non-equidistant interval lengths.

Before we discuss what can be done to address this stiffness in the context of our algorithm, we first consider what the algorithm requires in terms of numerical results and the integration process:

1. It is important to realise that we do not need to integrate until a steady state solution is obtained. The required integration time should only be long enough to observe all the relevant model dynamics. Figure 4.3 is an example of the dynamics of this stiff system, where state x_1 evolves much slower than the other 2 states. By increasing the integration time, we can ensure that all the system dynamics is observed. Figure 4.4 shows the dynamics of x_1 when t_N is increased from 0.5 to 10000. The dynamics in figure 4.4 is ideal for the analysis of a model even though $10^4 < 10^9$. Keep in mind that some states may already be in steady state at time $t = 0$, and so increasing the integration time will not result in an increase in observed dynamics.

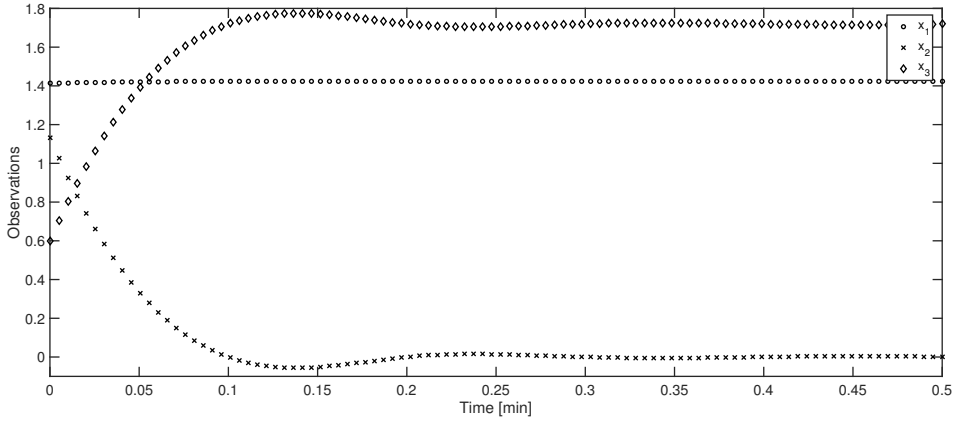


Figure 4.3: **Typical dynamics of the stiff Robertson problem over a short integration interval:** Over this short integration interval, the time evolution of state x_1 is much slower than that of the other model states.

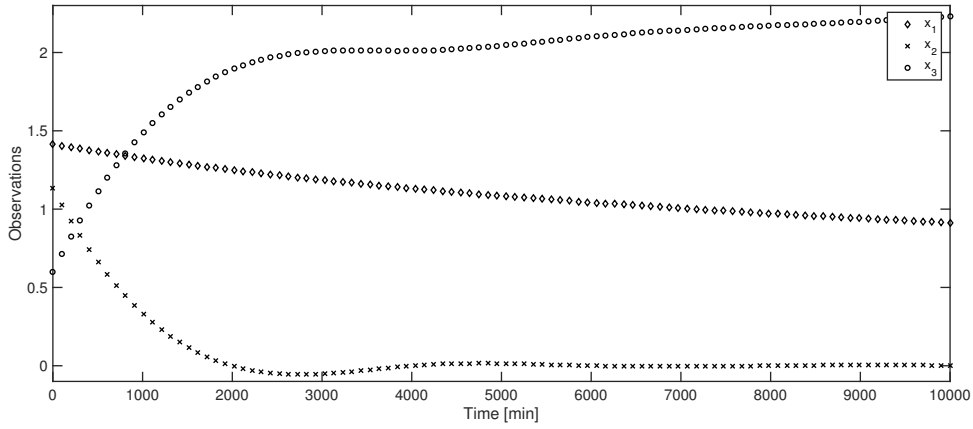


Figure 4.4: **Typical dynamics of the stiff Robertson problem over a long integration interval:** Integrating the model over a longer time interval reveals the model dynamics of all three model states. These dynamics are ideal for the identifiability analysis of a model.

2. The numerical integrator should compute the solution within seconds. Only then can the identifiability algorithm be used in applications where multiple iterations are required.

Practical tips:

1. First analyse the model using a set of randomly generated parameter and initial values. If numerical integration becomes computationally demanding, a set of optimised parameter values could be obtained by minimising an objective function that essentially is based on the condition number of the Jacobi matrix, $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}$.

The condition number of a matrix is the ratio of the largest and the smallest singular values. A value of 1 indicates a well conditioned system. This initial step will give the user a general impression of the structural identifiability of model his/her model.

2. If the user wishes to analyse the model for specific initial conditions, these values should be used in the second round of model analysis.

3. If no dynamics is observed over a period $[t_0, t_N]$, investigate whether there is really a steady state present or whether there are states with different time scales.

4. Select a numerical integrator suitable for the analysis of stiff ODE systems. ODE15s and ODE23s are both MATLAB-based methods suited to the analysis of such models.

5. Verify results symbolically.

4.3.3. NUMBER OF MEASURED SENSORS

In this section, the effect of the number of states/sensors measured on the gap size is studied. We use the JAK/STAT model as example, for which it is known that when states x_{10} and x_{11} are not measured, the model is structurally unidentifiable with both $x_{10}(0)$ and $x_{11}(0)$ not identifiable [9].

Pitfall: The gap size between the singular values may not be significant if a model is analysed measuring an output that contains too few of states/sensors.

Figure 4.5 shows the effect of the number of states/sensors measured on the gap size. It indicates that for small output sets, the gap size reduces to 0.62. Also notice that when we measure as little as 8 of the 29 possible states, a clear gap of 4 decades already emerges.

Practical tips: Due to our method's efficient computation times, some additional analyses can be done to determine key states/sensors whose omission affect the model's identifiability. For example, if a researcher can only measure states x_1 and x_{31}, and wishes to know whether the model is identifiable or not, the following strategy should be followed:

1. First analyse the model for the user defined output.

2. Apply the minimal output set (MOS) algorithm developed in chapter 2 to identify key states/sensors whose omission from the output will result in structural unidentifiability. In chapter 2, we discussed the random omission of states/sensors from an output, and we used the Bernoulli equation to compute the number of iterations required to identify indispensable sensors with a 99.5% probability.

3. Verify whether the states/sensors identified in the MOS algorithm, are present in the defined output. If they are not, the model will be classified as structurally unidentifiable, regardless of the gap size obtained in 1. Because the researcher in this example does not measure states x_{10} or x_{11}, the model is structurally unidentifiable despite a gap size of 0.62.

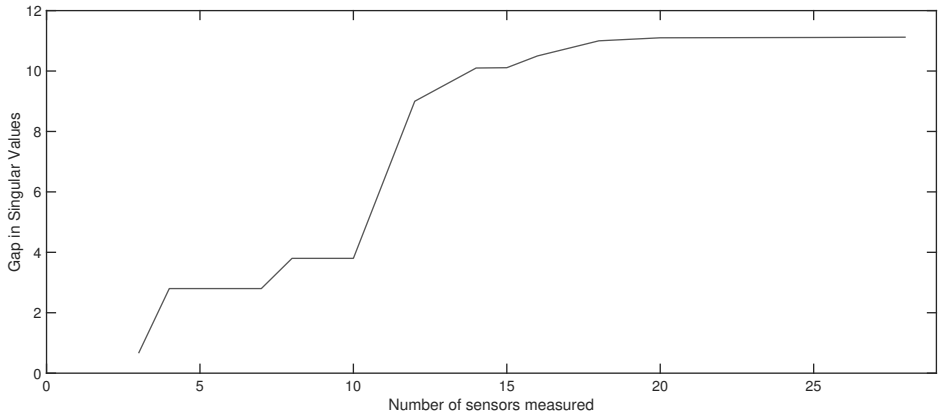


Figure 4.5: **Gap between the singular values as a function of the number of states/sensors measured (length of the output vector, y).** The gap size is reduced below our chosen threshold of 3 decades when only 2 or 3 states are measured. The full model has 31 states and we gradually include more states into the measured output. To force the model to be unidentifiable, we always omit states x_{10} and x_{11}. As shown in chapter 2, x_{31} is the only other state that influences structural identifiability results and accordingly, this state is included into all measured output sets so as to not introduce additional unidentifiable parameter sets.

4. To determine which parameters are unidentifiable measuring a small output, a model can subsequently be analysed using a concatenated sensitivity matrix. This will be discussed in section 4.7.

5. Results should be verified symbolically.

4.4. USER SPECIFIED FACTORS THAT CAN BE ADJUSTED

Given that our method relies on both integration and differentiation, the following adjustable factors will influence numerical results:

1. $\boldsymbol{\theta}$ - The parameter and initial values.

2. t_N - The integration time interval.

3. N - The number of discretized points in the time interval. $N = n_\theta$ indicates that the number of points is equal to the number of unknown parameters.

4. Integr - The integration method used.

5. $\boldsymbol{y}$ - The length and components of the model's output vector.

6. Tol - The numerical integration tolerance specified.

7. Norm – Whether or not a normalised sensitivity matrix is used. Note that normalisation is not possible if certain initial conditions are zero.

8. S/N - Whether the matrices in equation 4.4, $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}$ and $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\theta}}$, are calculated symbolically or numerically.

9. Exp - The number of parameter sets to analyse which is equivalent to the number of sensitivity matrices to vertically concatenate. For more on this topic, refer to section 4.7.

The adjustable factors are summarised in figure 4.6. In the following section, the sensitivity of the results to each of these factors is investigated for a number of very different examples. The metrics used to measure their influence on the numerical method are: 1) the *size of the gap* between the singular values (a large gap makes unidentifiability results easy to interpret. In this thesis we assume that a gap larger than 3 decades warrants a symbolic analysis) and, 2) numerical *integration times*.

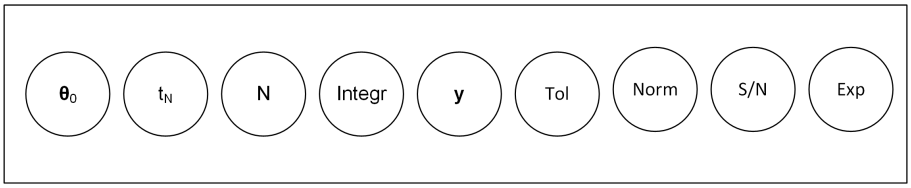


Figure 4.6: **Potential factors that will influence identifiability results.** These could affect both gap size and computation times.

A model is always first analysed using rule of thumb (R.O.T) settings, and these are:

Settings: $\boldsymbol{\theta} = \mathbf{0.5} + \boldsymbol{rand}[\mathbf{0,1}]$ (all parameter values are randomly generated in the interval [0.5, 1.5]), $\boldsymbol{t_N} = \mathbf{0.5}$ (the initial integration interval length), $\boldsymbol{N} = \boldsymbol{n_\theta}$ (the number of points on the integration interval equals the number of unknown parameters. This ensures that the sensitivity matrix is at least square when only 1 state/sensor is measured), $\boldsymbol{ODE45}$ (numerical integration method - this is a builtin MATLAB function), $\boldsymbol{Tol} = \mathbf{10^{-15}}$, (absolute tolerance of integration method), **Non normalised** (analysis using a non normalised sensitivity matrix), $\boldsymbol{S}$ ($\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}$ and $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\theta}}$ are generated symbolically, using MATLAB's symbolic toolbox. Alternatively, these matrices can be calculated numerically using the built-in function in MATLAB's Simbiology toolbox, which uses complex-step approximation to calculate sensitivities [10]), $\boldsymbol{Exp} = \mathbf{1}$ (no matrix concatenation and so a model is analysed for a single set of parameter and initial values).

In the following examples, the sensitivity of the results related to each of these factors is investigated.

4.5. EXAMPLES

EXAMPLE 1: JAK/STAT MODEL WITH 31 STATES

We start by analysing a relatively large ODE model, well-known to the identifiability community [9]. The aim is to identify key factors that influence both the gap size and computation times. This JAK/STAT model has 51 system parameters, 31 states and for this example, we choose to analyse its identifiability measuring 18 states: $\boldsymbol{y} = [x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{17}, x_{18}, x_{19}, x_{31}]$. Notice that states x_{10} and x_{11} are omitted from the measured output and we therefore know beforehand that the model is structurally unidentifiable [9].

RESULTS USING RULE OF THUMB SETTINGS

Settings: $\theta = 0.5 + rand[0,1]$, $t_N = 0.5$, $N = n_\theta$, $ODE45$, $Tol = 10^{-15}$, Non normalised, S, $Exp = 1$.

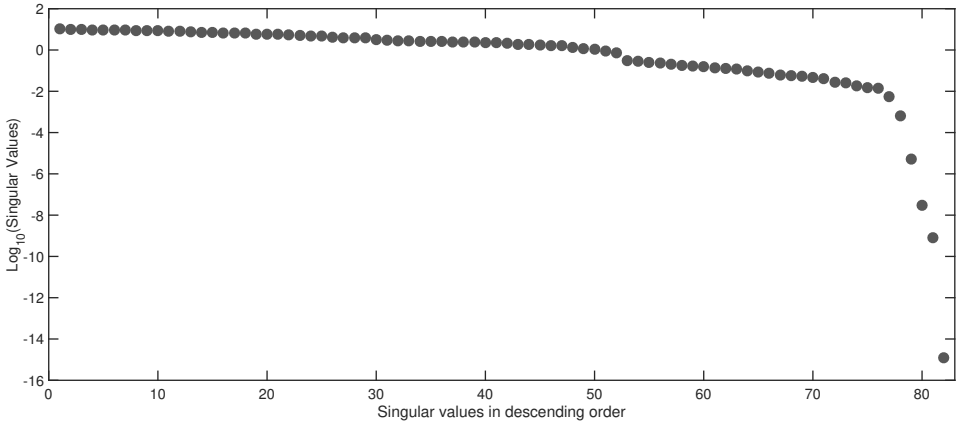


Figure 4.7: **JAK/STAT model:** Gap size of 6.34 is obtained using the rule of thumb settings. This indicates that the model is structurally unidentifiable when states x_{10} and x_{11} are not measured.

Using the R.O.T settings, the calculated gap size is 6.34 (figure 4.7) and time required for numerical integration is 0.66 seconds. With only 1 singular value beyond a gap of 3 decades, we expect that the 4 unidentifiable parameters are totally correlated. This result is confirmed by the symbolically computed nontrivial null-space: $\mathcal{N}\left(\frac{d\boldsymbol{G}}{d\boldsymbol{\theta}^{\boldsymbol{unid}}}(\boldsymbol{\theta})\right) = \{\theta_{14}/x_{11}(0), -\theta_{51}/x_{11}(0), x_{10}(0)/x_{11}(0), 1\}$, where $\boldsymbol{\theta}^{unid} = \{\theta_{14}, \theta_{51}, x_{10}(0), x_{11}(0)\}$. This initial analysis shows that the algorithm gives satisfactory results when using the R.O.T settings. Let us now take a closer look at the influence of individual factors on results.

1. THE EFFECT OF PARAMETER VALUES, $\boldsymbol{\theta}$

In this analysis, values for both system parameters and initial conditions are randomly generated from a uniform distribution in MATLAB. Here, we investigate the sensitivity of the gap size to these generated values in two different analyses. For convenience, we will refer to the entire set of system parameters and initial conditions as "parameter values". We first look at the influence of the parameter values on the gap size and then at the effect of the magnitude of difference between the parameter values.

Settings: $t_N = 0.5$, $N = n_\theta$, $ODE45$, $Tol = 10^{-15}$, Not normalised, S, $Exp = 1$.

In the first experiment, values are generated by adding a randomly chosen value between 0 and 1 to a constant value, which in this analysis, is systematically increased from 0.1 to 10. This is described by $(*) + rand[0,1]$, where $* = 0.1, \ldots, 10$. This implies that the maximum difference between individual parameters is 1.

Figure 4.8 reveals a slight increase in the gap size as the parameter values increase. These larger gap sizes do however come at a price, with the accompanying computation times increasing significantly (figure 4.9). A compromise must therefore be struck between the gap size and time. Importantly, the identifiability results are in principle

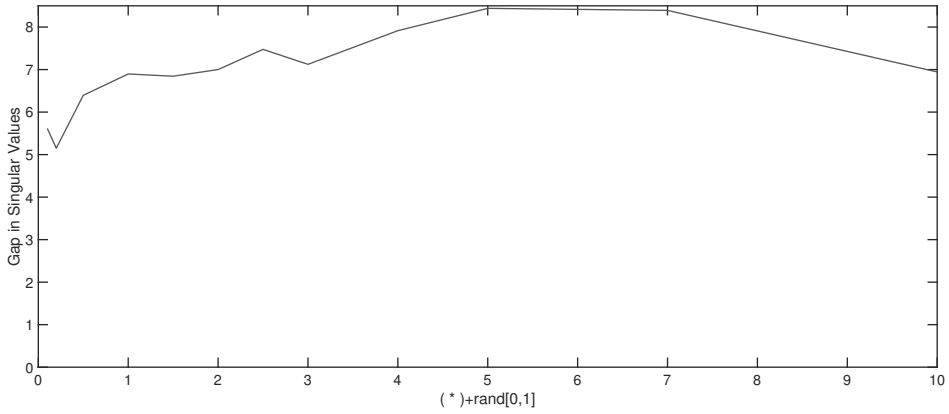


Figure 4.8: **JAK/STAT model:** Gap sizes calculated for randomly generated parameter values. Parameter values are calculated as $(*) + rand[0,1]$, where $*$ is systematically increased from 0.1 to 10. Analysed parameter sets therefore range from $[0.1, 1.1]$ to $[10, 11]$.

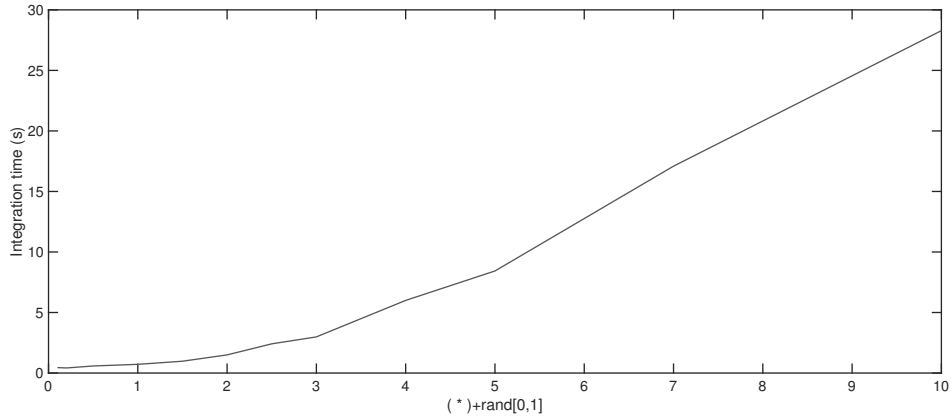


Figure 4.9: **JAK/STAT model:** Integration times associated with the randomly generated parameters values in figure 4.8. These are calculated as $(*) + rand[0,1]$, where $*$ is systematically increased from 0.1 to 10.

insensitive to parameter values. This is confirmed by a gap larger than 5 decades maintained for all parameter sets.

In the second analysis we systematically increase the span between the randomly generated values and add them to a constant value of 0.5. This is described by $(0.5) + rand[0,*]$, where $* = 1, \ldots, 10$. For example, when $* = 10$, the maximum difference between parameter values is 10. Figure 4.10 reveals no sensitivity to the span between parameters. We therefore use $0.5 + rand[0,1]$ as a R.O.T for generating values for the system parameters and initial conditions, since these lower values also reduce integration times.

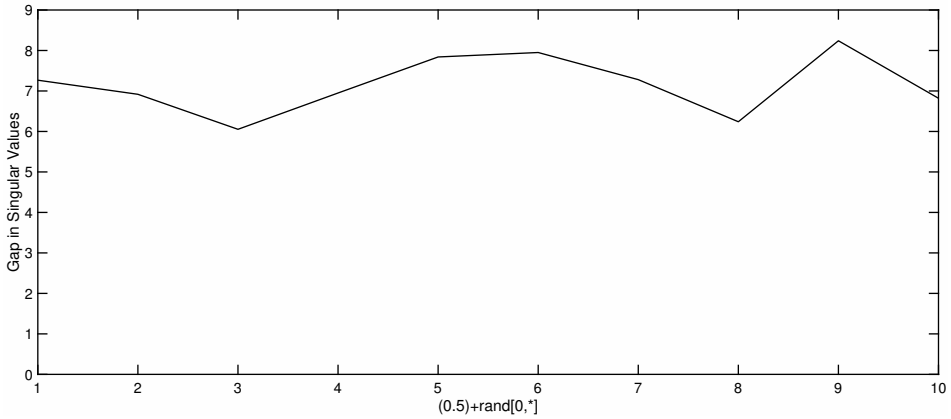


Figure 4.10: **JAK/STAT model:** Gap size for the values of randomly generated parameters. These are calculated as $(0.5) + rand[0, *]$, where $*$ is systematically increased from 1 to 10.

2. EFFECT OF THE CHOICE OF INTEGRATION INTERVAL, t_N

Settings: $\theta = 0.5 + rand[0,1]$, $N = n_\theta$, $ODE45$, $Tol = 10^{-15}$, Non normalised, S, $Exp = 1$.

In this section, we investigate the effect of the length of the integration interval on both gap size and computation time. Keep in mind that the integration time, t_N, should be long enough to allow for the observation of model dynamics. The suggested integration time for a set of randomly generated parameter values from the interval $[0.5, 1.5]$ is 0.087 (the largest eigenvalue of the Jacobi matrix is $\lambda_{Max} = 11.49$, and so $t_N = 1/11.49 = 0.087$). The system is nonstiff for these particular parameter values and suggests that the rule of thumb interval, $[0, 0.5]$, is sufficient.

Figures 4.11 and 4.12 show the relationship between t_N and the gap sizes and computation times, respectively. Despite the small theoretically required integration length of 0.087, results are sensitive to smaller times. As expected, there is an upper limit to the obtainable gap size as errors associated with longer integration times influence numerical accuracy. With $t_N = 0.5$, the calculated gap of 6.34 is sufficient and therefore we suggest using this as R.O.T.

3. EFFECT OF THE CHOICE OF INTEGRATION DISCRETIZATION, N

Settings: $\theta + 0.5 + rand[0,1]$, $t_N = 0.5$, $ODE45$, $Tol = 10^{-15}$, Non normalised, S, $Exp = 1$.

The lower limit of N is determined by the number of states/sensors in the output vector to ensure that the sensitivity matrix is at least square. Figure 4.13 shows that the number of grid points on the integration interval do not effect the size of the gap. Nor does it influence integration times. We suggest using $N = n_\theta$, the number of unknown parameters that need to be inferred from measured data.

4. EFFECT OF THE CHOICE OF INTEGRATION METHOD

Settings: $\theta + 0.5 + rand[0,1]$, $t_N = 0.5$, $N = N_\theta$, $Tol = 10^{-15}$, Non normalised, S, $Exp = 1$.

The relatively small required integration length that follows from the largest eigenvalue of the Jacobi matrix, $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}$, $t_N = \frac{1}{|\lambda_{Max}|} = 0.087$, suggests that the system is nonstiff for the

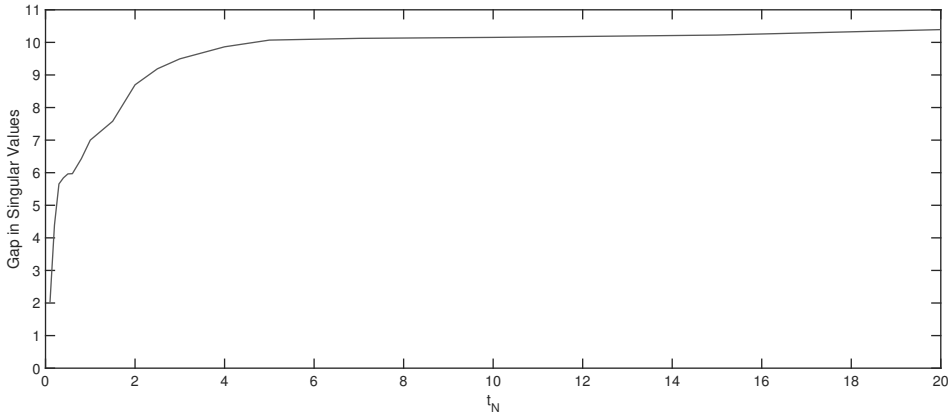


Figure 4.11: **JAK/STAT model:** Gap size obtained for the different integration times, t_N. For $t_N = 0.5$, a gap of 6.34 is calculated, which strongly suggests that this integration interval is sufficient in allowing for an adequate amount of observed model dynamics.

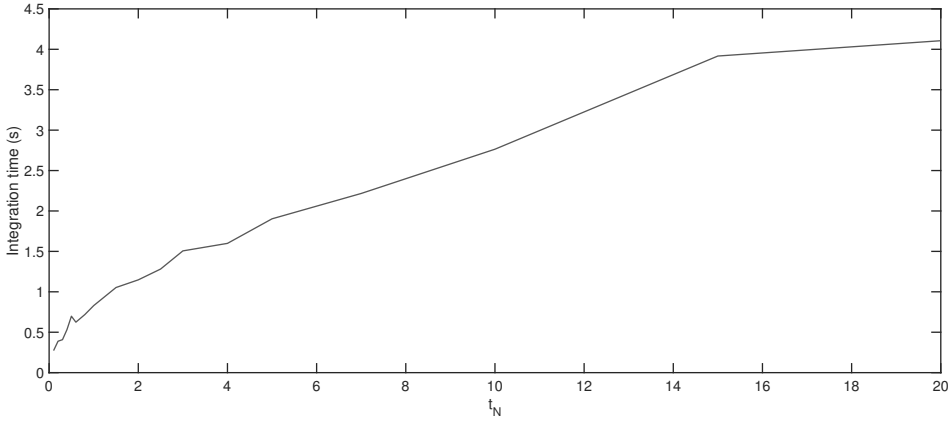


Figure 4.12: **JAK/STAT model:** Integration times for the different values of t_N. As t_N increases, there is almost a linear increase in the associated integration times.

randomised parameter values. Given the efficient integration times of approximately 0.6 seconds, ODE45 is used as the default integration method.

5. EFFECT OF THE MEASURED OUTPUT SIZE AND COMPOSITION, $\boldsymbol{y}$
Settings: $\theta = 0.5 + rand[0,1]$, $t_N = 0.5$, $N = n_\theta$, $ODE45$, $Tol = 10^{-15}$, Non normalised, S, $Exp = 1$.

Figure 4.5 on page 104 shows the effect of the length of the output vector on the calculated gap size. We therefore know that measuring more states/sensors usually results in sharper identifiability results. This is clearly an important factor.

In section 4.3.3 it was mentioned that performing the MOS algorithm can be very

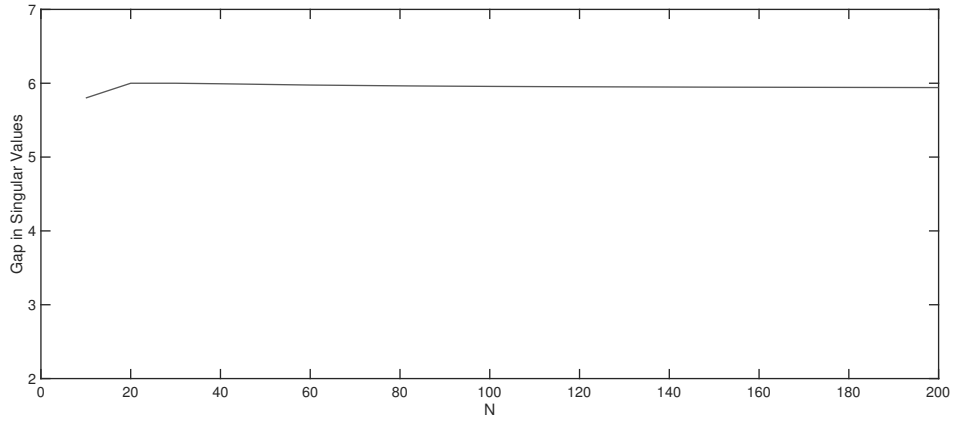


Figure 4.13: **JAK/STAT model:** Gap size obtained for different numbers of grid points, N, on the integration interval.

informative. Here, it can be used to limit the impact of this factor on results by systematically identifying important states/sensors who's omission result in unidentifiability.

6. EFFECT OF THE CHOICE OF INTEGRATION TOLERANCE
Settings: $\theta = 0.5 + rand[0,1]$, $t_N = 0.5$, $N = n_\theta$, $ODE45$, Non normalised, S, $Exp = 1$.

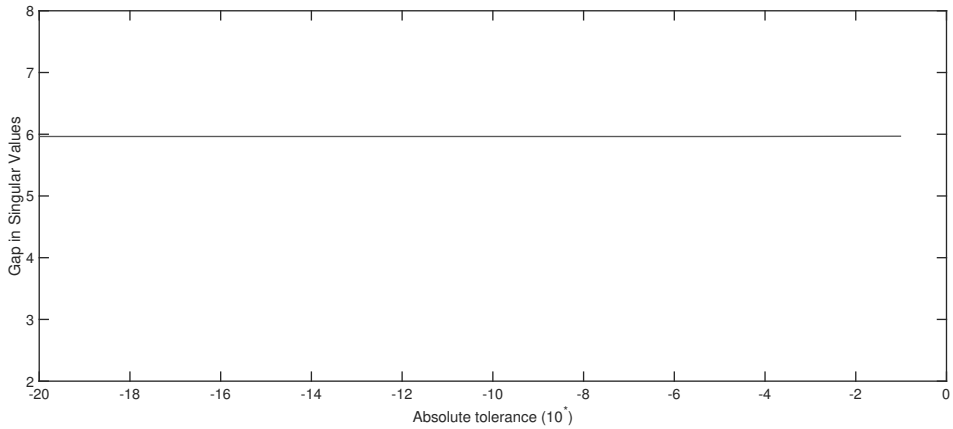


Figure 4.14: **JAK/STAT model:** Gap size obtained for different absolute tolerance values of the numerical integrator. Here, by substituting *, values range between 10^{-1} to 10^{-20}.

From figure 4.14 we know that for this example, this factor does not influence results and the R.O.T setting of 10^{-15} is used.

7. EFFECT OF USING A NORMALISED SENSITIVITY MATRIX
Settings: $\theta = 0.5 + rand[0,1]$, $t_N = 0.5$, $N = n_\theta$, $ODE45$, $Tol = 10^{-15}$, S, $Exp = 1$.

Table 4.1: Influence of normalising the sensitivity matrix on the gap size in the singular values

Type	Gap size
Normalised	6.0959
Non normalised	6.139

Normalising the sensitivity matrix does not improve results and so non-normalised matrices are used.

8. EFFECT OF GENERATING MATRICES $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}$ AND $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\theta}}$ SYMBOLICALLY OR NUMERICALLY
Settings: $\theta = 0.5 + rand[0,1]$, $t_N = 0.5$, $N = n_\theta$, $ODE45$, $Tol = 10^{-15}$, Not normalised, $Exp = 1$.

As stated earlier, these matrices can be calculated numerically using MATLAB's Simbiology toolbox. This is particularly useful for large ODE models, for which the use of symbolic calculations can be laborious. Here, we look at the effect on the gap sizes when calculating these matrices numerically as opposed to symbolically. From table 4.2 we conclude that computing $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}$ and $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\theta}}$ numerically, significantly affects the results. The user is therefore urged, if the size of the model permits it, to calculate these matrices symbolically.

Table 4.2: Influence of the method used to calculate matrices $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}$ and $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\theta}}$, on both integration times and the gap size.

Type	Time (s)	Gap size
Symbolic	0.6486	6.0959
Numerical	0.484987	1.8644

EXAMPLE 2: JAK/STAT MODEL WITH 14 STATES

This model has 14 states, 22 system parameters and 8 observed outputs [11].

RESULTS USING RULE OF THUMB SETTINGS

Settings: $\theta = 0.5 + rand[0,1]$, $t_N = 0.5$, $N = n_\theta$, $ODE45$, $Tol = 10^{-15}$, Non normalised, S, $Exp = 1$.

It has several zero initial conditions, $x_3(0) = x_5(0) = x_7(0) = x_9(0) = x_{10}(0) = x_{12}(0) = x_{13}(0) = x_{14}(0) = 0$. Unidentifiable results are indicated by a gap of 11.35 and numerical integration was completed in 1.39 seconds. For this model, R.O.T settings are sufficient and the numerical results are confirmed by the symbolically calculated nontrivial null-space, $\mathcal{N}\left(\frac{d\boldsymbol{G}}{d\boldsymbol{\theta}^{\boldsymbol{unid}}}(\boldsymbol{\theta})\right) = \{0,0,-\theta_{17}/\theta_{22},0,1\}, \{-\theta_{11}/\theta_{21},-\theta_{15}/\theta_{21},0,1,0\}$, where $\boldsymbol{\theta}^{unid} = \{\theta_{11}, \theta_{15},\theta_{17},\theta_{21},\theta_{22}\}$.

EXAMPLE 3: CHINESE HAMSTER MODEL

Let us now analyse a very large ODE model. It has 117 system parameters, 34 states and 13 observed states/sensors, $\boldsymbol{y} = [x_1,x_2,x_3,x_4,x_5,x_{11},x_{13},x_{15},x_{21},x_{27},x_{29},x_{30},x_{32}]$ [5,

12]. It does not have a predefined set of initial conditions and so values for both system parameters and initial conditions have to be chosen.

RESULTS USING RULE OF THUMB SETTINGS

Settings: $\theta = 0.5 + rand[0,1]$, $t_N = 0.5$, $N = n_\theta$, $ODE45$, $Tol = 10^{-15}$, Non normalised, S, $Exp = 1$.

For this example, not all rule of thumb settings apply since the model size does not allow for the symbolic calculation of the matrices, $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}$ and $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\theta}}$, in acceptable times. As a result, the settings are adjusted to allow for the numerical computation of these matrices. *Settings*: $\theta = 0.5 + rand[0,1]$, $t_N = 0.5$, $N = n_\theta$, $ODE45$, $Tol = 10^{-15}$, Non normalised, $\boldsymbol{N}$, $Exp = 1$.

The calculated gap size is 7.57, with numerical integration completed in 1.18 seconds (figure 4.15). These result were confirmed by the symbolically computed nontrivial null-space: $\mathcal{N}\left(\frac{d\boldsymbol{G}}{d\boldsymbol{\theta}^{\boldsymbol{unid}}}(\boldsymbol{\theta})\right) = \{1,1,0,0\}\{0,0,1,1\}$, where $\boldsymbol{\theta}^{unid} = \{\theta_{47}, \theta_{48}, \theta_{55}, \theta_{57}\}$.

In the following sections we investigate the influence of different adjustable factors when using the numerical approach implemented in the Simbiology toolbox.

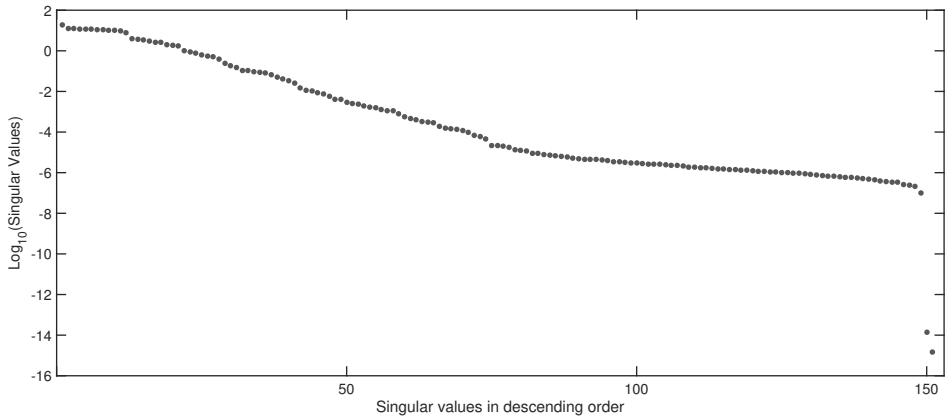


Figure 4.15: **Chinese Hamster model:** A gap size of 7.57 is obtained by numerically calculating matrices $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}$ and $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\theta}}$.

1. THE EFFECT OF THE SET OF PARAMETER VALUES, $\boldsymbol{\theta}$

Settings: $t_N = 0.5$, $N = n_\theta$, $ODE45$, $Tol = 10^{-15}$, Non normalised, N, $Exp = 1$.

Table 4.3 shows the gap sizes computed for 2 different parameter vectors, $\boldsymbol{\theta}^1$ and $\boldsymbol{\theta}^2$ respectively. As parameter values increase, the gap size decreases and the integration times increase. For values in the interval $[4,5]$ and larger, the gap disappears. Given that for R.O.T settings, large gap sizes are obtained, parameter values generated in the interval $[0.5,1.5]$ are sufficient.

From the results in table 4.3 we can conclude that the method is sensitive to parameter values for this example. Later we will see that vertically concatenating individual

sensitivity matrices, each analysed for a different parameter trajectory considerably improves results and makes the algorithm more robust to parameter values.

Table 4.3: Influence of randomised parameter values on the gap size in the singular values. Values are systematically increased by substituting * with the values in column 1.

	$\boldsymbol{\theta}^1$	$\boldsymbol{\theta}^2$	
$(*)+rand[0,1]$	Gap	Gap	Integr time (s)
$*=0.05$	Only 1 singular value	6.11	1.18
$*=0.1$	4.7449	Only 1 singular value	1.07
$*=0.3$	4.0386	5.28	1.19
$*=0.5$	7.8166	4.744	1.73
$*=0.7$	3.5399	Only 1 singular value	1.18
$*=1$	3.3171	Only 1 singular value	1.18
$*=2$	0.873	1.455	3.36
$*=3$	1.31	1.143	7

2. EFFECT OF THE CHOICE OF INTEGRATION INTERVAL, t_N

Settings: $\theta = 0.5 + rand[0,1]$, $N = n_\theta$, $ODE45$, $Tol = 10^{-15}$, Non normalised, N, $Exp = 1$.

The method is sensitive to the length of the integration interval, with longer times introducing numerical round-off errors that obscure the gap. Figure 4.16 shows results for integration intervals between 0.2 and 2. These all result in a clear gap. This range of integration intervals investigated, allows for the observation of sufficient dynamics and so, for this example the R.O.T setting of $t_N = 0.5$ suffices.

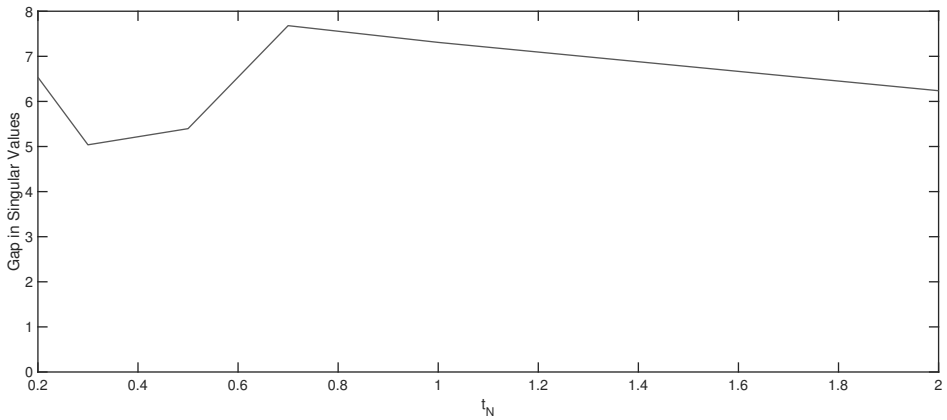


Figure 4.16: **Chinese Hamster model:** Gap size obtained for the different integration intervals, t_N. Intervals between 0.2 and 2, allow for the observation of sufficient model dynamics and keep numerical errors within bounds to allow for the observation of a gap larger than 5 decades.

3. EFFECT OF THE CHOICE OF INTEGRATION DISCRETIZATION, N

Settings: $\theta = 0.5 + rand[0,1]$, $t_N = 0.5$, $ODE45$, $Tol = 10^{-15}$, Non normalised, N, $Exp = 1$.

Figure 4.17 reveals the insignificant role the number of grid points on the integration interval, $[0, 0.5]$, has on numerical accuracy. This corresponds to the result observed for the JAK/STAT model in example 1 and so the method is robust to different discretization scenarios.

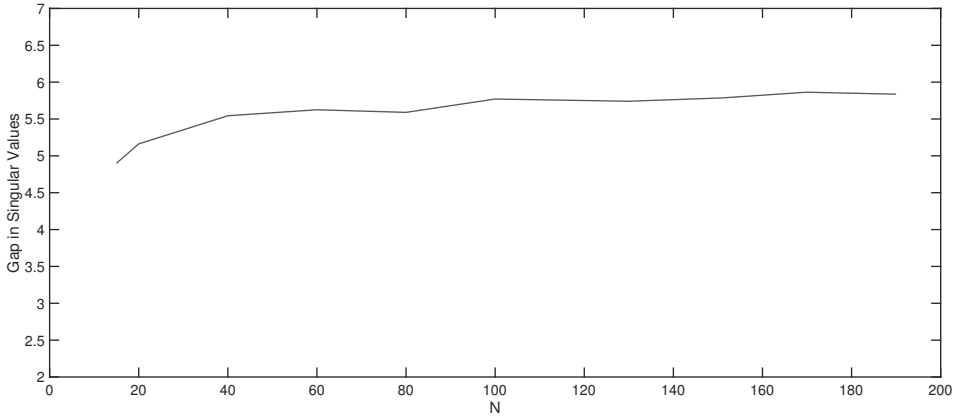


Figure 4.17: **Chinese Hamster model:** Gap size obtained for different number of grid points, N, on the integration interval. The increased sampling rate of a particular set of measured states/sensors does not significantly affect numerical results.

4

4. EFFECT OF THE CHOICE OF INTEGRATION METHOD

Settings: $\theta = 0.5 + rand[0,1]$, $t_N = 0.5$, $N = n_\theta$, $Tol = 10^{-15}$, Non normalised, N, $Exp = 1$.

Using ODE45, numerical integration requires more than a second for completion. In a bid to improve efficiency, the model is also analysed using ODE15s. This integration method does not improve calculation times and we conclude that using $ODE45$ is sufficient.

Table 4.4: Influence of the integration method used on integration time and the gap size.

Type	Time (s)	Gap size
ODE45	1.157	5.7861
ODE15s	1.087	5.7861

5. EFFECT OF THE MEASURED OUTPUT SIZE, $\boldsymbol{y}$

Settings: $\theta = 0.5 + rand[0,1]$, $t_N = 0.5$, $N = n_\theta$, $ODE45$, $Tol = 10^{-15}$, Non normalised, N, $Exp = 1$.

We see from figure 4.18 that unidentifiability is only detected if the number of states/sensors is larger than 5. As we saw with the JAK/STAT model in figure 4.18, there is a maximum obtainable gap size where measuring additional states/sensors does not contribute to numerical accuracy.

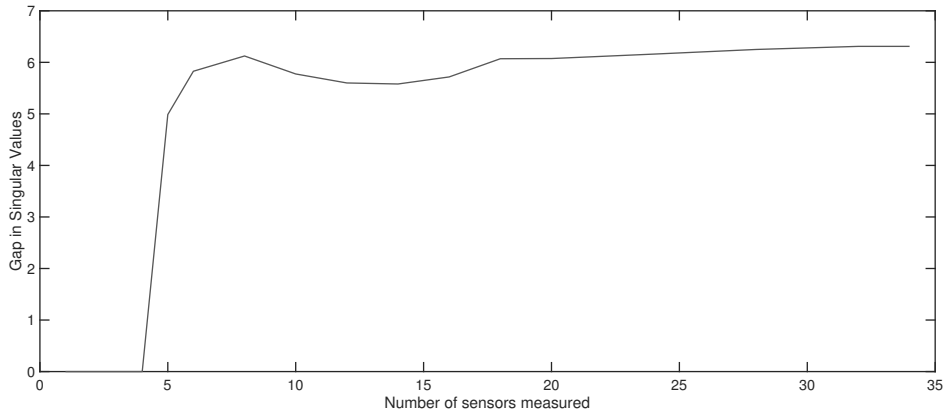


Figure 4.18: **Chinese Hamster model:** Gap size obtained as a function of the number of states/sensors measured.

6. EFFECT OF THE CHOICE OF INTEGRATION TOLERANCE
Settings: $\theta = 0.5 + rand[0,1]$, $t_N = 0.5$, $N = n_\theta$, $ODE45$, Non normalised, N, $Exp = 1$.

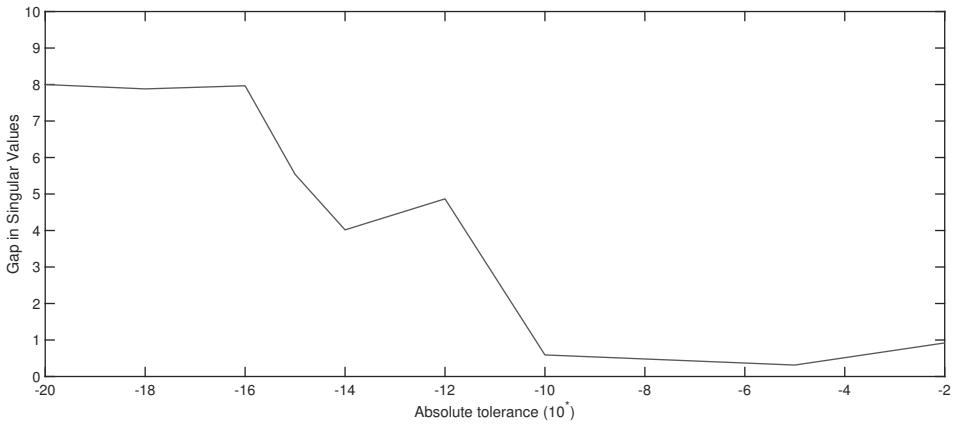


Figure 4.19: **Chinese model results:** Gap size obtained for different absolute tolerance values of the numerical integrator. The tolerance is varied in the range $[10^{-20}, 10^{-2}]$.

Tolerances are typically adjusted to limit the local discretization error. More specifically, the absolute tolerance determines the accuracy when solutions approach zero and so we adjust this value in a bid to obtain sharper results. From figure 4.19 we see that setting the absolute tolerance to 10^{-15} is sufficient for generating a clear gap, and that as this tolerance is relaxed, the gap becomes insignificant.

7. EFFECT OF THE CHOICE OF USING A NORMALISED SENSITIVITY MATRIX
Settings: $\theta = 0.5 + rand[0,1]$, $t_N = 0.5$, $N = n_\theta$, $ODE45$, $Tol = 10^{-15}$, N, $Exp = 1$.

Table 4.5: Influence of normalising the sensitivity matrix on the gap size

Type	Gap size
Normalised	5.5396
Non normalised	5.0358

In table 4.5, we see no significant difference in the calculated gap sizes and so we opt to not normalise values, thereby enabling the analysis of model with initial conditions of zero.

EXAMPLE 4: NOVAK TYSON MODEL

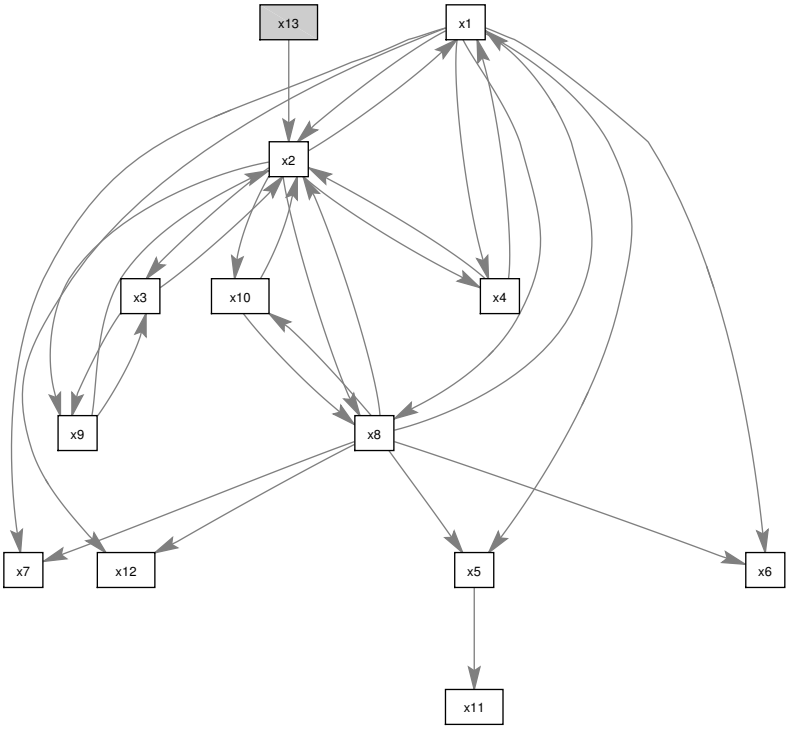


Figure 4.20: **Novak Tyson model structure:** Applying Liu *et. al.*'s root strongly connected components principle [13], the directed graph does not suggest that state x_{13} should be included into the model's minimal output sets. However, our analysis reveals that if this state is not measured, the model is structurally unidentifiable.

This model has 39 system parameters and 13 states [14]. The initial conditions of all these states are assumed to be unknown. Analysing its directed graph in figure 4.20 and applying the root strongly connected components principle for Liu *et. al.* [13], it is apparent that states x_6, x_7, x_{11} and x_{12} should be measured to ensure the model's structural

identifiability. Here, the model is analysed to see if additional important states/sensors can be identified and to determine whether standard settings are sufficient. We find that state x_{13} should also be measured.

RESULTS USING RULE OF THUMB SETTINGS

Settings: $\theta = 0.5 + rand[0,1]$, $t_N = 0.5$, $N = n_\theta$, $ODE45$, $Tol = 10^{-15}$, Non normalised, S, $Exp = 1$.

Implementing our minimal output set algorithm to identify essential sensor sets, the first round of analysis entails the omission of single states/sensors from the measured output. Measuring the output $\boldsymbol{y} = [x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}]$, thus omitting state x_{13}, a gap size of 9.1 was calculated using R.O.T settings (figure 4.21). This numerical result was confirmed by the symbolically computed nontrivial null-space: $\mathcal{N}\left(\frac{d\boldsymbol{G}}{d\boldsymbol{\theta}^{\boldsymbol{unid}}}(\boldsymbol{\theta})\right) = \{-\theta_9/x_{13}(0), 1\}$, where $\boldsymbol{\theta}^{unid} = \{\theta_9, x_{13}(0)\}$.

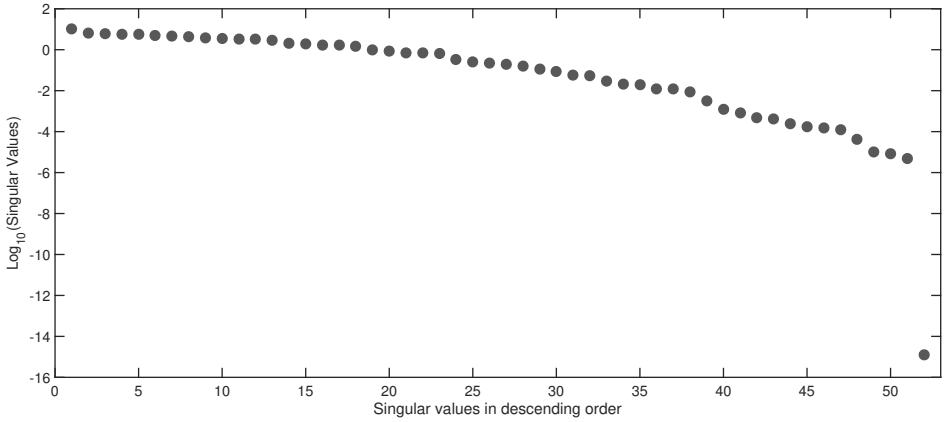


Figure 4.21: **Novak Tyson model:** If all states but x_{13} are measured, a gap of 9.1 is obtained using the rule of thumb settings.

EXAMPLE 5: MODEL WITH 20 STATES

We now investigate a model analysed by Saccomani *et. al.* in 2010 [15]. It has 20 states and 22 system parameters. We assume that all initial conditions are unknown.

RESULTS USING RULE OF THUMB SETTINGS

Settings: $\theta = 0.5 + rand[0,1]$, $t_N = 0.5$, $N = n_\theta$, $ODE45$, $Tol = 10^{-15}$, Non normalised, S, $Exp = 1$.

We implement our minimal output set algorithm to identify potential important states/sensors. Not measuring x_{20} and using R.O.T settings, the calculated gap of 12.9 is shown in figure 4.22. This result was confirmed by the symbolically computed nontrivial null-space: $\mathcal{N}\left(\frac{d\boldsymbol{G}}{d\boldsymbol{\theta}^{\boldsymbol{unid}}}(\boldsymbol{\theta})\right) = \{1,0\}\{0,1\}$, where $\boldsymbol{\theta}^{unid} = \{\theta_{22}, x_{20}(0)\}$.

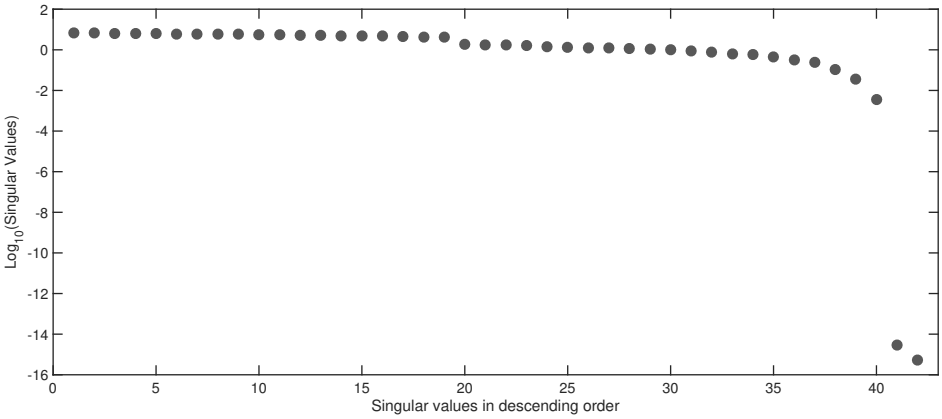


Figure 4.22: **Model with 20 states:** If all states but x_{20} are measured, a gap of 12.9 is obtained using the rule of thumb settings.

EXAMPLE 6: POLLUTION MODEL

Here, we analyse a stiff ODE system with 20 states and 25 systems parameters [8, 16]. The following states have zero-initial conditions: $x_1(0), x_3(0), x_5(0), x_6(0), x_{10}(0), x_{11}(0), x_{12}(0), x_{13}(0), x_{14}(0), x_{15}(0), x_{16}(0), x_{18}(0)$, $x_{19}(0)$ and $x_{20}(0)$.

RESULTS USING RULE OF THUMB SETTINGS

Settings: $\theta = 0.5 + rand[0,1]$, $t_N = 0.5$, $N = n_\theta$, $ODE45$, $Tol = 10^{-15}$, Non normalised, S, $Exp = 1$.

Measuring as output all states except x_{20}, a gap of 10.7 is calculated. This result was confirmed by the symbolically computed nontrivial null-space: $\mathcal{N}\left(\frac{d\boldsymbol{G}}{d\boldsymbol{\theta}^{\boldsymbol{unid}}}(\boldsymbol{\theta})\right) = \{1\}$, where $\boldsymbol{\theta}^{unid} = \{x_{12}(0)\}$. By generating parameter values in the interval, $[0.5, 1.5]$, numerical integration was completed in 0.57 seconds using $ODE45$. Sufficient model dynamics are observed on the interval $[0, 0.5]$ and finally, the sensitivity matrix could not be normalised due to the fact that some of the observed initial conditions are zero.

EXAMPLE 7: LUNG CANCER MODEL

The final model analysed in this chapter contains 21 states and 54 system parameters [17, 18].

RESULTS USING RULE OF THUMB SETTINGS

Settings: $\theta = 0.5 + rand[0,1]$, $t_N = 0.5$, $N = n_\theta$, $ODE45$, $Tol = 10^{-15}$, Non normalised, S, $Exp = 1$.

From the directed graph of the model, one can observe that states x_1 and x_9 are root strongly connected components and therefore need to be measured to ensure the model's structural identifiability. A description of the model and its directed graph is given in chapter 3. Measuring all but states x_1 and x_9, no clear unidentifiability result is found analysing the model using R.O.T settings. We now investigate the effect of individual factors on this result.

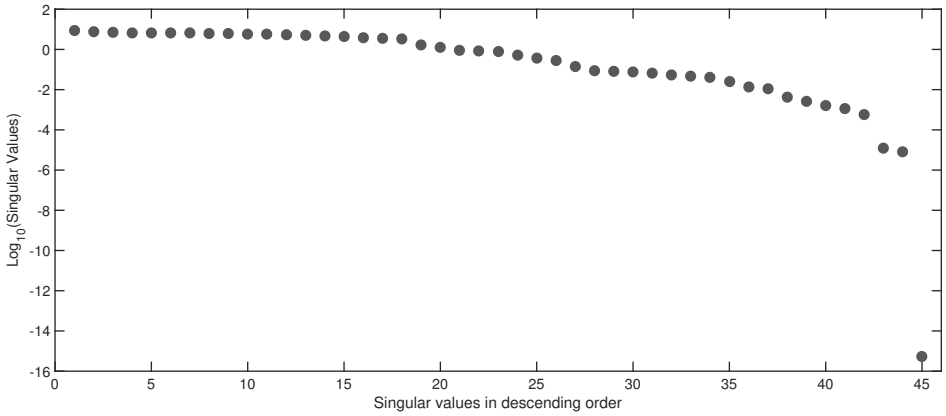


Figure 4.23: **Pollution model results:** Measuring all but state x_{20}, a gap size of 10.7 was obtained using the rule of thumb settings.

1. THE EFFECT OF THE SET OF PARAMETER VALUES, $\boldsymbol{\theta}$
Settings: $t_N = 0.5$, $N = n_\theta$, $ODE45$, $Tol = 10^{-15}$, Non normalised, S, $Exp = 1$.

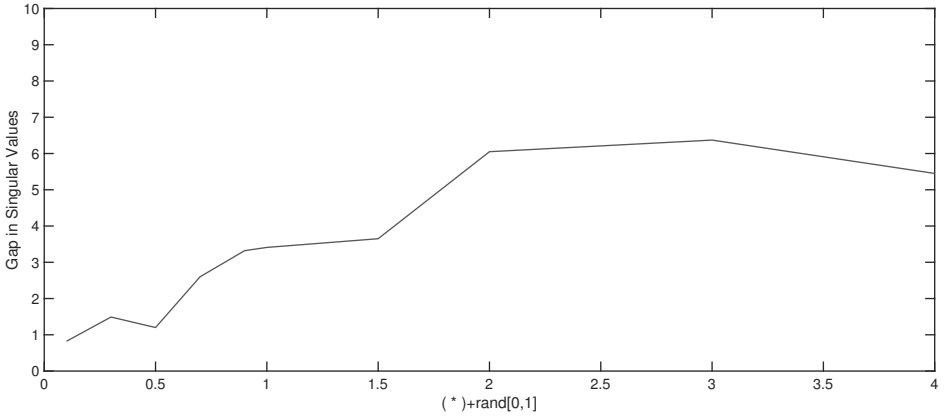


Figure 4.24: **Lung cancer model:** Gap size as a function of randomly generated parameter values, $(*) + rand[0,1]$, where $*$ ranges from 0.1 to 4.

Figure 4.24 reveals that a gap emerges as parameter values increase. This shows the method's sensitivity to parameter values. There are two singular values beyond this gap and this is confirmed by the symbolically computed nontrivial null-space: $\mathcal{N}\left(\frac{d\boldsymbol{G}}{d\boldsymbol{\theta}^{\boldsymbol{unid}}}(\boldsymbol{\theta})\right) = \{0, -\theta_{17}/x_9(0), -\theta_{20}/x_9(0), 0, 0, 1\}\{-\theta_6/x_1(0), 0, 0, -2\theta_{23}/x_1(0), 1, 0\}$, where $\boldsymbol{\theta}^{unid} = \{\theta_6, \theta_{17}, \theta_{20}, \theta_{23}, x_1(0), x_9(0)\}$. Ideally, one wishes to implement standardised settings, and so we continue with our analysis by keeping $\theta = 0.5 + rand[0,1]$ and search for additional factors that might sharpen results.

2. EFFECT OF THE CHOICE OF INTEGRATION INTERVAL, t_N

Settings: $\theta = 0.5 + rand[0,1]$, $N = n_\theta$, $ODE45$, $Tol = 10^{-15}$, Non normalised, S, $Exp = 1$.

From figure 4.25 we see that the integration time interval plays a vital role in whether or not a gap is evident. Obtaining sharp numerical results therefore require longer integration intervals to capture more dynamics of the system. We saw previously that these sharper results come at a price as is evident in the accompanying required integration times in figure 4.26. We suggest using slightly longer integration intervals when the span between the eigenvalues of $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}$ is large, keeping figure 4.26 and the potential numerical round-off errors associated with larger integration intervals in mind.

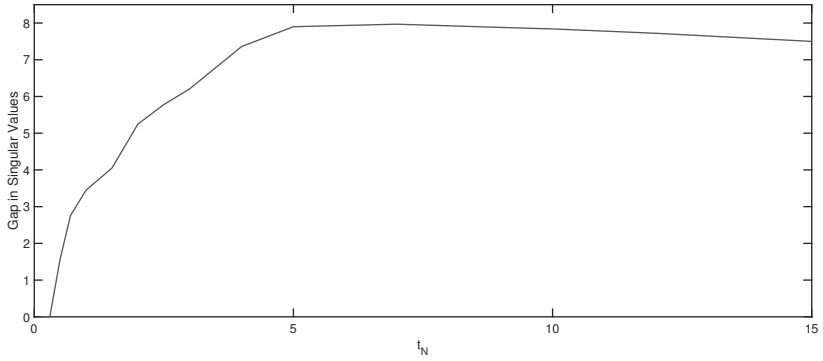


Figure 4.25: **Lung cancer model:** Gap size obtained for different integration interval times, t_N.

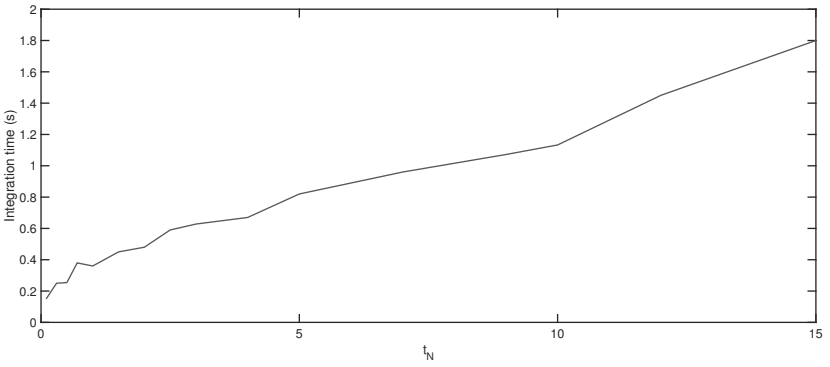


Figure 4.26: **Lung cancer model:** Required integration times for the different integration intervals, t_N.

3. EFFECT OF THE CHOICE OF INTEGRATION DISCRETIZATION, N

Settings: $\theta = 0.5 + rand[0,1]$, $\boldsymbol{t_N = 5}$, $ODE45$, $Tol = 10^{-15}$, Non normalised, S, $Exp = 1$.

For this model, the method is insensitive to the choice of the discretization, N. From figure 4.27 we do however witness a lower threshold and therefore we use $N = n_\theta$ as R.O.T setting.

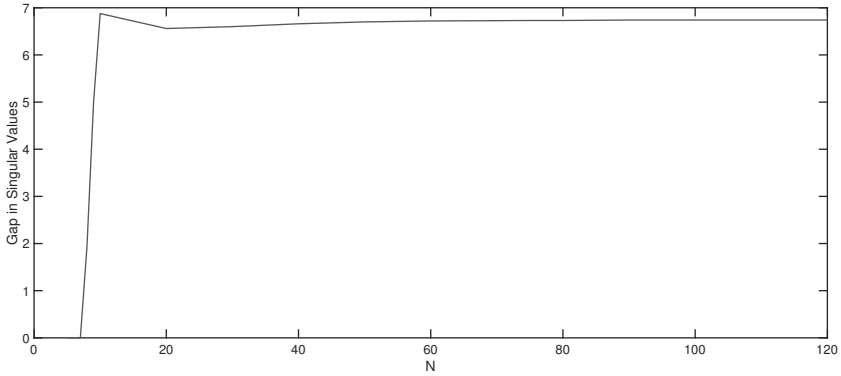


Figure 4.27: **Lung cancer model:** Gap size obtained for different number of grid points on the integration interval, N.

4. EFFECT OF THE CHOICE OF INTEGRATION METHOD

Settings: $\theta = 0.5 + rand[0,1]$, $\boldsymbol{t_N} = \mathbf{5}$, $N = n_\theta$, $Tol = 10^{-15}$, Non normalised, S, $Exp = 1$.

Table 4.6: Influence of integration method used on integration time and the gap size.

Type	Time (s)	Gap size
ODE45	0.58	6.466
ODE15s	13.66	6.466

Table 4.6 suggests that there is no difference in gap sizes for the different integration methods and that $ODE45$ is superior in terms of computation times. Accordingly, $ODE45$ is the preferred integration method.

5. EFFECT OF THE MEASURED OUTPUT SIZE, $\boldsymbol{y}$

Settings: $\theta = 0.5 + rand[0,1]$, $\boldsymbol{t_N} = \mathbf{5}$, $N = n_\theta$, $ODE45$, $Tol = 10^{-15}$, Non normalised, S, $Exp = 1$.

From figure 4.28 it is apparent that a gap already becomes visible when as little as five states/sensors are measured.

6. EFFECT OF THE CHOICE OF INTEGRATION TOLERANCE

Settings: $\theta = 0.5 + rand[0,1]$, $\boldsymbol{t_N} = \mathbf{5}$, $N = n_\theta$, $ODE45$, Non normalised, S, $Exp = 1$.

The choice of the absolute tolerance for the numerical integration method hardly affects the gap size (figure 4.29).

7. EFFECT OF THE CHOICE OF USING A NORMALISED SENSITIVITY MATRIX

Settings: $\theta = 0.5 + rand[0,1]$, $\boldsymbol{t_N} = \mathbf{5}$, $N = n_\theta$, $ODE45$, $Tol = 10^{-15}$, S, $Exp = 1$.

Table 4.7 shows that normalising the sensitivity matrix does not improve results.

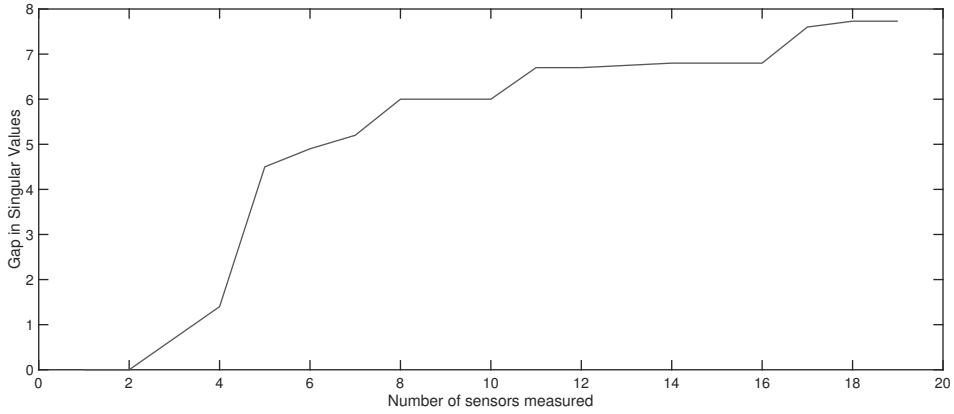


Figure 4.28: **Lung cancer model:** Gap size obtained as a function of the number of states/sensors in the output vector. In this example states, x_1 and x_9 are omitted from these measured output.

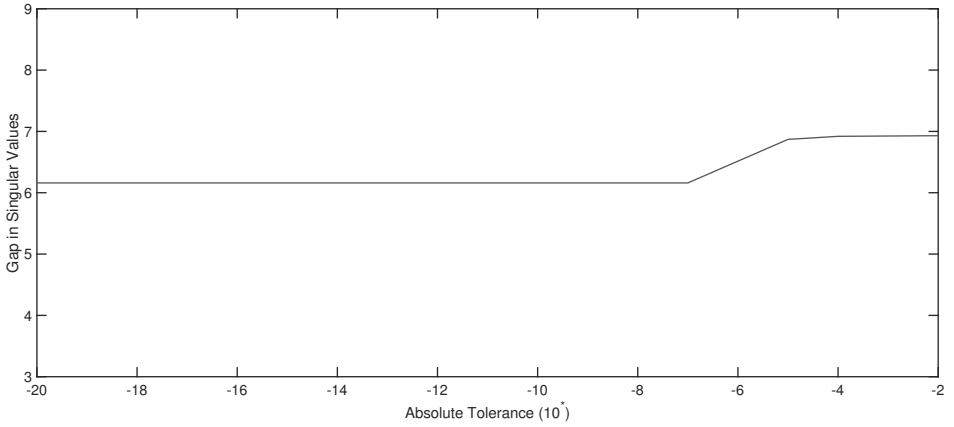


Figure 4.29: **Lung cancer results:** Gap size obtained for different absolute tolerance values ranging from 10^{-20} to 10^{-2}.

Table 4.7: Influence of normalising the sensitivity matrix on the gap size.

Type	Gap size
Normalised	6.448
Non normalised	6.47

8. EFFECT OF OF GENERATING MATRICES $\frac{\partial f}{\partial x}$ AND $\frac{\partial f}{\partial \theta}$ SYMBOLICALLY OR NUMERICALLY

Settings: $\theta = 0.5 + rand[0,1]$, $\boldsymbol{t_N = 5}$, $N = n_\theta$, $ODE45$, $Tol = 10^{-15}$, Non normalised, $Exp = 1$.

Table 4.8 shows that for this example, symbolically calculated matrices, $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial \theta}$, offer superior results.

Table 4.8: Influence of method used to calculate matrices $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial \theta}$, on integration times and the gap size.

Type	Time (s)	Gap size
Symbolic	0.58	6.47
Numerical	0.14	0.9

4.6. SUMMARY OF RESULTS

Table 4.9: Summary of the structural identifiability results for the models analysed in this chapter. ✓ indicates that the model is correctly classified as unidentifiable using the R.O.T settings. The most important factors when R.O.T settings do not apply, indicated by ×, are given in the last column.

Model	States	Parms	Output	Gap	ROT	Different from R.O.T
Long Jak/Stat	31	51	18	6.34	✓	
Short Jak/Stat	14	22	8	11.35	✓	
Chinese Hamster	34	117	13	7.57	×	N
Novak Tyson	13	39	12	9.1	✓	
High dimensional	20	22	19	12.9	✓	
Pollution	20	25	19	10.7	✓	
Lung Cancer	21	54	19	6.47	×	$\boldsymbol{\theta}$ or t_N

4.7. MATRIX CONCATENATION

Table 4.9 reveals that for most examples, the rules of thumb are sufficient in offering clear unidentifiability results, with all the gaps larger than 6 decades. Two of the most influential factors are the *parameter values* and the *length of the output vector*. In this section, we introduce a concept aimed at reducing their influence on numerical results. It involves the vertical concatenation of different sensitivity matrices, each evaluated for a particular parameter set, corresponding to a specific trajectory in the state space. The result is a large sensitivity matrix as defined in equation 4.8.

We start this discussion by illustrating the advantage of this multiple trajectory approach. Consider the JAK/STAT model with 31 states as example. For the measured output, $\boldsymbol{y} = [x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{17}, x_{18}, x_{19}, x_{31}]$, (so states x_{10} and x_{11} are not measured). Table 4.10 compares the gap size obtained analysing a single parameter trajectory, versus the gap computed using multiple parameter trajectories. Concatenation significantly improves the sharpness of the numerical results.

Table 4.10: Effect of analysing a concatenated sensitivity matrix on the gap size. No concatenation - only 1 parameter trajectory analysed. 10 matrices concatenated - 10 different parameter trajectories analysed.

	Gap size
No concatenation	6.34
Concatenating (10 matrices)	9.85

We know from Anguelova *et. al.* [9] that simultaneously omitting states x_{10} and x_{11} from the model's output results in the unidentifiability of 4 parameters. In the context

of the sensitivity matrix, this unidentifiability can be described as the linear dependence between 4 columns of this matrix. For this example, this relation reads as:

$$\frac{\theta_{14}}{x_{11}(0)}\frac{\partial \boldsymbol{y}}{\partial \theta_{14}} - \frac{\theta_{51}}{x_{11}(0)}\frac{\partial \boldsymbol{y}}{\partial \theta_{51}} + \frac{x_{10}(0)}{x_{11}(0)}\frac{\partial \boldsymbol{y}}{\partial x_{10}(0)} + \frac{\partial \boldsymbol{y}}{\partial x_{11}(0)} = 0. \tag{4.21}$$

In general, equation 4.21 can be written as:

$$c_1\frac{\partial \boldsymbol{y}}{\partial \theta_{14}} + c_2\frac{\partial \boldsymbol{y}}{\partial \theta_{51}} + c_3\frac{\partial \boldsymbol{y}}{\partial x_{10}(0)} + c_4\frac{\partial \boldsymbol{y}}{\partial x_{11}(0)} = 0. \tag{4.22}$$

The main idea behind the vertical concatenation of multiple sensitivity matrices is increasing the number of rows of a sensitivity matrix. More specifically, increasing the number of rows by adding matrices, each evaluated in different parameter trajectories, decreases the numerical error of the calculated singular values. This should however be done in such a way so as to *preserve the values of the coefficients* c_1, c_2, c_3 *and* c_4 in 4.22. This implies that all parameter values related to these 4 columns must be kept constant, and that all other parameter values may be changed.

Matrix concatenation comprises the following steps:

1. First observe potential unidentifiable parameters. For this example, these are θ_{14}, $\theta_{51}, x_{10}(0)$ and $x_{11}(0)$. This can be done by analysing the model with the minimal output set algorithm.
2. Repeat the identifiability method k times, each time calculating a different sensitivity matrix. Importantly, the values of the potentially unidentifiable parameters identified in step 1, $\theta_{14}, \theta_{51}, x_{10}(0)$ and $x_{11}(0)$, are kept constant to preserve the constants defined in equation 4.21.
3. Vertically concatenate the k different sensitivity matrices, $\boldsymbol{S}^i$, to generate a matrix $\boldsymbol{S_{all}}$ as defined in 4.8.
4. Perform an SVD analysis on the concatenated matrix.

Table 4.10 indicated the potential role concatenation could play in making the method robust to parameter values. Let us now consider its potential benefit in making the algorithm less susceptible to output vector sizes. Figure 4.30 shows the results obtained when measuring less than half of the states measured for the results in table 4.10. Measuring only seven states, $\boldsymbol{y} = [x_1, x_2, x_3, x_4, x_5, x_6, x_{31}]$, and evaluating one set of parameter values reveals no gap. However, a gap can be seen when evaluating a concatenated sensitivity matrix, comprising of ten different sensitivity matrices in figure 4.31. This indicates that by implementing concatenation, we can reduce the algorithm's sensitivity to the length of the output vector.

Finally, we perform an analysis to determine the influence of the number of concatenated matrices on the gap size. Figure 4.32 shows the gap size as a function of the number of vertically concatenated matrices, measuring two different output sets. For one of these sets, we reduce the number of states measured to 2 and so $\boldsymbol{y}_1 = [x_1, x_{31}]$. This is indicated in blue in figure 4.32. The second contains the seven states measured in figures 4.30 and 4.31, $\boldsymbol{y}_2 = [x_1, x_2, x_3, x_4, x_5, x_6, x_{31}]$ and is indicated in red. Figure 4.32 shows that concatenating as little as 4 matrices results in clear gap sizes measuring either $\boldsymbol{y}_1$ or $\boldsymbol{y}_2$. We also see that there is a numerical threshold of 9 decades.

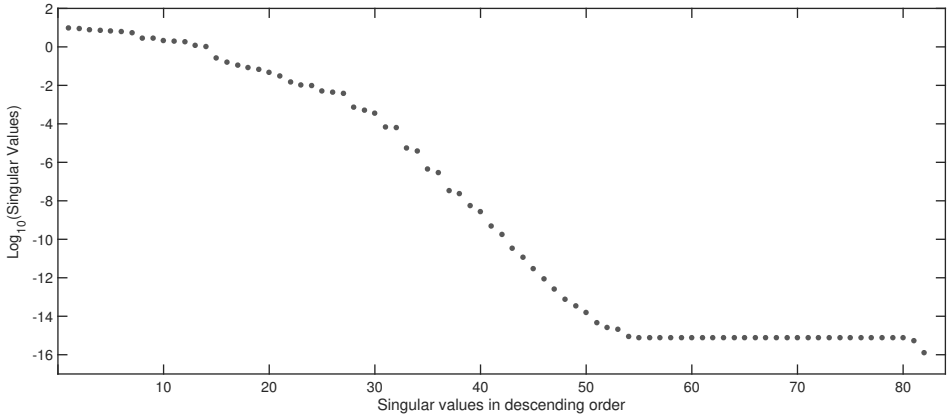


Figure 4.30: **Large Jak/Stat:** Singular values calculated whilst measuring the reduced output, $\boldsymbol{y} = [x_1, x_2, x_3, x_4, x_5, x_6, x_{31}]$, and evaluating the model for only one set of parameter values.

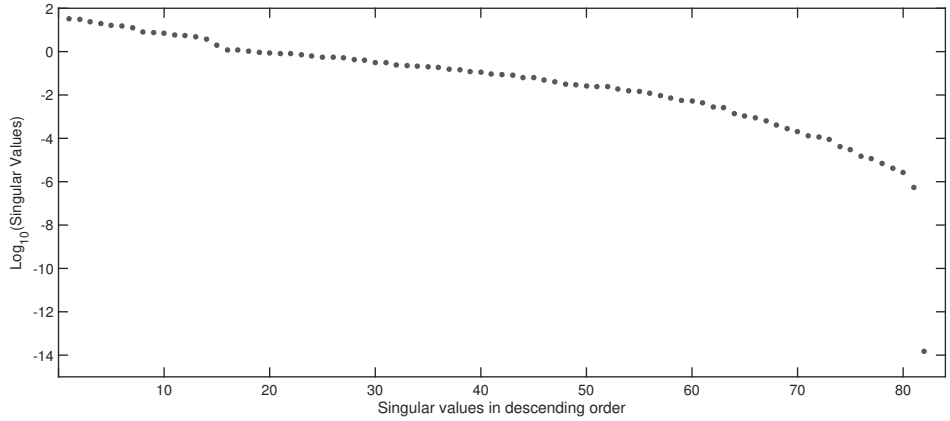


Figure 4.31: **Large Jak/Stat:** Gap size obtained measuring reduced output, $\boldsymbol{y} = [x_1, x_2, x_3, x_4, x_5, x_6, x_{31}]$, implementing rule of thumb settings and concatenating 10 sensitivity matrices. This is in stark contrast to the set of singular values generated using the non concatenated matrix in figure 4.30.

4.8. CONCLUSION

This chapter we investigated the robustness of the identifiability method studied in this thesis. Due to its numerical nature, several questions regarding the influence of factors such as parameter values, the rate of sampling, and the number of sensors measured had to be answered. Here, we methodically analysed the effect of each of these factors on 2 quantitative metrics, the *gap size* in the singular values and the *time* required for numerical integration. The motivation for writing this chapter was to gain a better understanding of the influence of individual factors on results, with the intended purpose of developing a structural identifiability software package.

To summarise, the most influential factors are *parameter values* and the *length and*

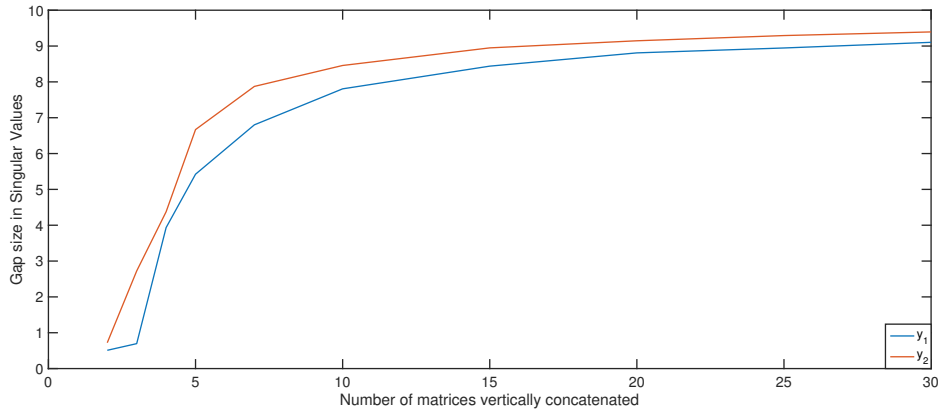


Figure 4.32: **Large Jak/Stat :** Gap size obtained for different numbers of concatenated matrices. The measured outputs are, $\boldsymbol{y_1} = [x_1, x_{31}]$ (blue) and $\boldsymbol{y_2} = [x_1, x_2, x_3, x_4, x_5, x_6, x_{31}]$ (red), respectively. The gap size increases as the number of concatenated matrices increases. Also note that he difference in the number of matrices required increases as the length of the output vector decreases.

composition of the output vector. For certain models, the length of the integration interval, t_N, may also be influential. Because the output vector is often defined my the modeller, we looked for potential ways to make our algorithm less sensitive to both parameter values and the number of measured sensors. We found that by vertically concatenating different sensitivity matrices, this sensitivity can be mitigated.
Having analysed several models, we propose the following rule of thumb settings:

- Generate values for both systems parameters and initial conditions in the interval $[0.5, 1.5]$ - by choosing all these values all in this narrow range, one avoids potential scaling problems,
- Use $ODE45$ as numerical integrator,
- Set the absolute numerical tolerance to 10^{-15},
- For most models, if parameter values are generated on the interval $[0.5, 1.5]$, there is no need to normalise the sensitivity matrix,
- Take as integration interval $[0, 0.5]$,
- The number of points on this interval should equal the number of unknown parameters,
- Supply the numerical integrator with symbolically generated matrices, $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}$ and $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\theta}}$, if possible.

After quantifying the sensitivity of results using numerous well-known models, we conclude that using our method and implementing standard settings is sufficient for most models. The only exceptions being the Chinese Hamster model, where the size of

the model necessitates the numerical calculation of matrices, $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial \theta}$, and the Lung cancer model, where longer integration times are required. To that end, it might be useful to consider the span between the eigenvalues of the Jacobi matrix, $\frac{\partial f}{\partial x}$, as these could indicate that longer integration intervals are required.

Given the increase in the number of large and complex ODE systems in a field such as systems biology, there is a need for fast and user friendly methods, capable of analysing these models. From the previous chapters, we already know that our method is efficient in terms of computation times. Here, we pinned down user settings that make this method reliable and we also indicated how these settings could be adjusted if needed.

REFERENCES

[1] N. Evans, S. Cheung, and J. Yates, *Structural identifiability for mathematical pharmacology: models of myelosuppression,* J Pharmacokinet Pharmacodyn **45**, 70 (2018).

[2] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th ed. (The Johns Hopkins University Press, 2013).

[3] G. Quintana-Ortí and E. S. Quintana-Ortí, *Parallel codes for computing the numerical rank,* Linear Algebra and its Applications **275-276**, 451 (1998).

[4] H. Pohjanpalo, *Systems identifiability based on the power series expansion of the solution,* Mathematical Biosciences **41**, 21 (1978).

[5] A. F. Villaverde, A. Barreiro, and A. Papachristodoulou, *Structural identifiability of dynamic systems biology models,* PLOS Computational Biology **20**, 1 (2016).

[6] L. Michaelis and M. Menten, *Die kinetik der invertinwirkung,* Biochemische Zeitschrift , 333 (1913).

[7] K. A. Johnson and R. S. Goody, *The original michaelis constant: Translation of the 1913 michaelis-menten paper,* Biochemistry **39**, 8264–8269 (2011).

[8] Dipartimento di Matematica, University of Bari, *Test set for initial value problem solvers,* `http://www.dm.uniba.it/~testset` (2012), accessed: 2018-06-15.

[9] M. Anguelova, J. Karlsson, and M. Jirstrand, *Minimal output sets for identifiability,* Mathematical Biosciences **239**, 139 (2012).

[10] J. Martins, I. Kroo, and J. Alonso, *An automated method for sensitivity analysis using complex variables,* in *38th Aerospace Sciences Meeting and Exhibit*, https://arc.aiaa.org/doi/pdf/10.2514/6.2000-689 .

[11] A. Raue, J. Karlsson, M. P. Saccomani, M. Jirstrand, and J. Timmer, *Comparison of approaches for parameter identifiability analysis of biological systems,* Bioinformatics **30**, 1440–1448 (2014).

[12] D. Joubert, J. Stigter, and J. Molenaar, *Determining minimal output sets that ensure structural identifiability,* PLoS One **13**, e0207334 (2018).

[13] Y. Liu, J. Slotine, and B. A.L., *Observability of complex systems,* Proc Natl Acad Sci **10**, 2460 (2013).

[14] C. Letellier, I. Sendiña Nadal, E. Bianco-Martinez, and M. Baptista, *A symbolic network-based nonlinear theory for dynamical systems observability,* Sci Rep **8**, 1 (2018).

[15] M. P. Saccomani, S. Audoly, G. Bellu, and L. D'Angió, *Examples of testing global identifiability of biological and biomedical models with the daisy software,* Computers in Biology and Medicine **40**, 402–407 (2010).

[16] F. Mazzia, J. R.Cash, and K. Soetaert, *A test set for stiff initial value problem solvers in the open source software r: Package detestset,* Journal of Computational and Applied Mathematics **236**, 4119 (2012).

[17] F. Bianconi, E. Baldelli, V. Ludovini, L. Crinó, A. Flacco, and P. Valigi, *Computational model of egfr and igf1r pathways in lung cancer: a systems biology approach for translational oncology,* Biotechnol Adv **Jan-Feb**, 142 (2012).

[18] F. Bianconi, E. Baldelli, V. Ludovini, L. Crinó, A. Flacco, and P. Valigi, *Egfr and Igf1r pathway in lung cancer,* (2012).

5

ASSESSING THE ROLE OF INITIAL CONDITIONS IN THE LOCAL STRUCTURAL IDENTIFIABILITY OF LARGE NONLINEAR DYNAMICAL MODELS

Dominique JOUBERT, Hans STIGTER, Jaap MOLENAAR

It happens frequently in the global identifiability applications that the property holds only generically, i.e. except for a "thin" set of initial conditions. In these situations the system is (incorrectly but forgivably) nevertheless declared to be (global) identifiable, excluding certain subsets of initial states.
(Saccomani, Audoly, D'Angi'o, 2003)

Based on: D. Joubert, J.D. Stigter and J. Molenaar, *Assessing the role of initial conditions in the local structural identifiability of large nonlinear dynamical models* - (under review)

ABSTRACT

MANY papers have been written on the topic of structural identifiability. Collectively, these contribute to the narrative that stresses the importance of this *a priori* analysis in the model development process. However, in many of these papers the story ends with a structurally unidentifiable model. This leaves a researcher, unfamiliar with the underlying theory, with no clear strategy on how to address this potential problem. In this chapter, we continue on this journey by identifying the source of a model's unidentifiability. It is well-understood that certain sets of initial conditions may result in local structural unidentifiability. We show that our algorithm is capable of detecting problematic initial conditions that, if changed, would reinstate a model's structural identifiability. This chapter comprises 6 examples, with the first, a well-known 2-state model chosen for illustrative reasons. The second example contains an input function. Example 3 illustrates that a model may posses multiple sets of problematic initial conditions and shows that ideally, these values need to be chosen so that they result in informative dynamics. In this example, the problematic initial values result in steady state conditions. In the fourth example, the importance of informative dynamics in further stressed and in examples 5 and 6 we identify the problematic initial conditions of relatively large systems to show that our approach is capable of doing this in an efficient way.

5.1. INTRODUCTION

Mathematical models can be used to describe features of a system. These models may contain unknown parameters that cannot be measured directly and therefore need to be estimated from experimental data. Whilst estimating the values of parameters, model developers often encounter numerous practical difficulties, including ill-conditioned matrices, large computer memory requirements, and insufficient time and money to perform the required number of experiments to develop full mechanistic models [1]. In this chapter, we focus on one of the major challenges in parameter estimation, the issue of identifiability. In particular, we look at the role of initial conditions in identifiability.

Parameter identifiability answers the question of whether or not it is possible to estimate *unique* parameter values from the collected input-output data. Factors such as the sensors measured, the model's structure and measurement errors all play a role. There are two types of identifiability. The first is based on the structure of the model and is know as structural or *a priori* identifiability. This is a binary property and assumes that all experimental measurements are perfect. It is a prerequisite for the second type, i.e. practical identifiability, which describes the ability to estimate parameters from observed data. It takes both the amount and the quality of the data into account [2].

One major aspect of structural identifiability is that it can be analysed prior to the experimental phase and therefore offers a unique opportunity for the preliminary design of experiments. More concisely, the conditions that result in the structural unidentifiability of a model can be identified and addressed before any expenses on wasteful experiments are incurred. To emphasise the importance of this assessment, we quote (Marc Russel Birtwistle, 2008) "We conclude that poorly designed experiments are not only wasteful, but can also be harmful to parameter identifiability." These conditions include: 1) which sensors should at least be measured to ensure identifiability (addressed in chapter 2), and 2) which initial conditions should be avoided, if possible, to ensure a model's structural identifiability. The latter issue is addressed in this chapter and offers a modeller the possibility to make his/her model structurally identifiable by simply choosing alternative initial conditions.

THE ROLE OF INITIAL CONDITIONS

Before we can discus the potential influence of initial conditions on a model's structural identifiability, we first need to define some classical system-theoretic properties of dynamical models. In particular, *controllability* and *reachability*. These topics are important in the context of this discussion since they have often been mentioned in publications on the loss of structural identifiability for special sets of initial conditions.

The problem of reachability is to find the set of all final states at a time t_N, denoted as $\boldsymbol{x}(t_N)$, that can be *reached* from a given initial state $\boldsymbol{x}(0)$, in a finite time $[0, t_N]$. In contrast, the controllability problem is to find the set of all initial states, $\boldsymbol{x}(0)$, that can be *driven* to a final fixed state, $\boldsymbol{x}(t_N)$, over the finite interval $[0, t_N]$. For nonlinear systems, reaching a particular state is also known as *accessibility* [3].

Developing dynamic models not only requires the formulation of differential and/or algebraic equations, but also the specification of realistic initial and/or boundary conditions. It is well known that a model's structural identifiability can be affected by the values of initial conditions [3–5]. Therefore, when a system evolves from a "problematic

set of initial conditions", it may be impossible to estimate certain of its parameters.

Denis-Vidal and co-authors alluded to the importance of initial conditions in the identifiability analysis of uncontrolled models [4]. Saccomani *et. al.* confirmed the importance of initial conditions by looking at their role in controlled models [5]. In their paper, the authors look at the role of accessibility/reachability in a model's structural identifiability. They state that it happens frequently in global identifiability analyses that the property only holds generically, i.e. except for a "thin" set of initial conditions. In these situations the system is (incorrectly but forgivably) declared to be (global) identifiable [5].

In their 2017 paper, Villaverde and Banga list the methods capable of detecting these problematic initial conditions, and suggest a comprehensive analysis strategy based on a differential geometry approach. The methods that can detect the local loss of structural identifiability for specific initial conditions include [3]:

1. Exact Arithmetic Rank method (EAR) - Proposed by Karlsson *et. al.* [6]. This method allows for the specification of initial conditions and uses these values in an efficient numerical procedure. This approach only works for rational systems.

2. DAISY - This differential algebra method is also designed for rational systems, and allows for the specification and analysis of initial conditions [7].

3. STRIKE-GOLDD - Meant for local structural identifiability analysis, it adopts a differential geometry approach to symbolically evaluate the Observability Condition and entails the successive computation of Lie derivatives [8]. It can also test particular initial conditions by changing the generic x in the Observability-Identifiability matrix to a set of particular initial conditions, x_0 [3].

To summarise, initial conditions form an inherent part of system's structure and may therefore influence its structural identifiability [4]. Accordingly, *it is important to take initial conditions into consideration when analysing the identifiability of nonlinear systems.*

In section 5.3 we show that our method can be added to the above mentioned list of methods capable of detecting the local structural unidentifiability of models evaluated for specific sets of initial conditions. Moreover, our method is unique since it can detect problematic values of large ODE systems within very short computation times. Experimental researchers can therefore make realistic choices on how to go about reinstating a model's identifiability, taking experimental limitations into account.

5.2. THEORY AND METHOD

MODEL DEFINITION

We begin with the definition of a typical ordinary differential equation model. These models often describe mass balances and can be very detailed, containing numerous model states and vast numbers of unknown parameters. Such dynamic models can in

most cases be written in the standard state-space form [9]:

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{f}_0(\boldsymbol{x}(t), \boldsymbol{\theta}) + \sum_{i=1}^{k} u_i(t)\boldsymbol{f}_i(\boldsymbol{x}(t), \boldsymbol{\theta}), \tag{5.1}$$

$$\boldsymbol{x}(0) = \boldsymbol{x_0}, \tag{5.2}$$

$$\boldsymbol{y}(t) = \boldsymbol{h}(\boldsymbol{x}(t), \boldsymbol{\theta}). \tag{5.3}$$

The state vector, $\boldsymbol{x}$, evolves on a manifold V in R^n. Functions $\boldsymbol{f}_i$, where $i = 0, \ldots, k$, and $\boldsymbol{h}$ are assumed to be analytical on V and C^∞ functions and so, their partial derivatives of any order exist and are continuous [10]. A model's input functions are contained in vector $\boldsymbol{u}(t) \equiv \{u_1, \ldots, u_k\}$. State variables are contained in vector $\boldsymbol{x}(t)$ ($\dim(\boldsymbol{x}) = n$), system parameters in vector $\boldsymbol{\theta}$ ($\dim(\boldsymbol{\theta}) = p$) and measured model outputs in vector $\boldsymbol{y}(t)$ ($\dim(\boldsymbol{y}) = m$). Initial values of the model states may also be unknown and in such cases, the initial condition vector may be parameterised through some additional parameters that then become part of the identification problem. The resulting unknown parameter vector, $\boldsymbol{\theta}$, then has $\dim(\boldsymbol{\theta}) = p + n$ [11].

LOCAL STRUCTURAL IDENTIFIABILITY ANALYSIS

Structural identifiability analyses assume noise-free data. The identifiability method implemented here uses the sensitivity of model outputs with respect to individual model parameters [12, 13]. The numerically obtained sensitivity results are subsequently used in symbolic structural identifiability calculations, and the combination of these approaches enables us to analyse large ODE models. Sensitivities are calculated from the following equations:

$$\frac{d}{dt}\left(\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\theta}}\right) = \frac{\partial \boldsymbol{f}_0}{\partial \boldsymbol{x}}\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\theta}} + \frac{\partial \boldsymbol{f}_0}{\partial \boldsymbol{\theta}} + \sum_{i=1}^{k}\left(\frac{\partial \boldsymbol{f}_i}{\partial \boldsymbol{x}}\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\theta}} + \frac{\partial \boldsymbol{f}_i}{\partial \boldsymbol{\theta}}\right)u_i, \tag{5.4}$$

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{\theta}} = \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}}\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\theta}} + \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{\theta}}. \tag{5.5}$$

One obtains the matrix function, $\partial \boldsymbol{y}/\partial \boldsymbol{\theta}$, as function of time by integrating equations 5.1 and 5.4 and substituting the solution into 5.5. By calculating these sensitivities at discrete time points on an interval $[t_0, \ldots, t_N]$, for specific parameter and initial values, one can construct a sensitivity matrix, $\boldsymbol{S}$. If any of the initial values of model states are unknown, their identifiability can easily be assessed by regarding them as additional parameters. In such cases, $\boldsymbol{S}$ has up to $p + n$ columns, each related to a specific parameter, $\theta_i, i = 1, \ldots, p + n$. The sensitivity matrix calculated for a single set of parameter and initial values is:

$$\boldsymbol{S} = \begin{pmatrix} \frac{\partial y_1}{\partial \theta_1}(t_0) & \dots & \frac{\partial y_1}{\partial \theta_{p+n}}(t_0) \\ \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial \theta_1}(t_0) & \dots & \frac{\partial y_m}{\partial \theta_{p+n}}(t_0) \\ \vdots & & \vdots \\ \frac{\partial y_1}{\partial \theta_1}(t_N) & \dots & \frac{\partial y_1}{\partial \theta_{p+n}}(t_N) \\ \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial \theta_1}(t_N) & \dots & \frac{\partial y_m}{\partial \theta_{p+n}}(t_N) \end{pmatrix}. \tag{5.6}$$

As was mentioned in chapter 4, the sensitivity matrix can also be normalised, generating the matrix $\boldsymbol{S}_{norm}$. Alternatively, individual sensitivity matrices, calculated for different sets of parameter and initial values, can be vertically concatenated. A full ranked matrix $\boldsymbol{S}$, is a sufficient condition for local structural identifiability [14, 15]. Rank deficiency of $\boldsymbol{S}$ can be attributed to two factors: 1) a model output may be insensitive to a specific parameter and in this instance, all the entries in the sensitivity matrix pertaining to this parameter are zero and the parameter is classified as unidentifiable, and 2) a model output may be sensitive to a particular parameter, but this sensitivity is counteracted by the sensitivity of the output to one or more other parameters. As a result, different columns of the sensitivity matrix are linearly dependent and this implies that these parameters are totally correlated and unidentifiable [16]. We determine the rank of the sensitivity matrix numerically using SVD, expressed as:

$$\boldsymbol{S} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T. \tag{5.7}$$

If $\boldsymbol{S}$ has $p+n$ columns, matrix $\boldsymbol{\Sigma}$ will have $p+n$ singular values on its diagonal and these are arranged in descending order. The rank of $\boldsymbol{S}$ is the number of nonzero singular values and, conversely, rank-deficiency is indicated by the presence of zero-valued singular values [17]. Due to numerical rounding errors, singular values are seldom exactly zero and so one uses as practical definition: Zero-valued singular values are values that fall beyond a distinct gap in the spectrum of singular values [18]. Once possible unidentifiability, based on the presence of zero-valued singular values has been established, unidentifiable parameters are recognised as the nonzero entries in the columns of the matrix $\boldsymbol{V}$, related to these vanishing singular values. Both the singular values and the unidentifiable parameters can graphically be illustrated in an easy to interpret identifiability signature [13].

SYMBOLIC STRUCTURAL IDENTIFIABILITY ANALYSIS

Next, the numerical results are verified using symbolic calculations. This requires the symbolic calculation of a so-called Jacobi matrix. The computational demand often associated with computing this matrix is reduced by utilising the preceding numerical results. Concretely, one only has to compute derivatives of the Lie derivatives with respect to the parameters that were suggested to be unidentifiable from the SVD analysis.

These are combined in a set, $\boldsymbol{\theta}^{\boldsymbol{unid}}$, and can contain both system parameters and initial conditions. We subsequently use the rank condition for local structural identifiability presented by Tunali and Tarn [19].

The Jacobi matrix of a model with no control input can be computed using Lie derivatives. A Lie derivative, mathematically defined as $\mathscr{L}_{\boldsymbol{f}_0}\boldsymbol{h}$, is the directional derivative of the smooth function, $\boldsymbol{h}$, with respect to the drift vector field, $\boldsymbol{f}_0$, which describes the model dynamics and is defined as:

$$\mathscr{L}_{\boldsymbol{f}_0}\boldsymbol{h} = \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}}\boldsymbol{f}_0. \tag{5.8}$$

Successive Lie derivatives are computed as:

$$\mathscr{L}^i_{\boldsymbol{f}_0}\boldsymbol{h} = \frac{\partial \mathscr{L}^{i-1}_{\boldsymbol{f}_0}\boldsymbol{h}}{\partial \boldsymbol{x}}\boldsymbol{f}_0. \tag{5.9}$$

We use a symbolic algebra package, Kwatny's ProPac add-on for Mathematica, to calculate the Lie derivatives [11]. In a generating series expansion, successive Lie derivatives of the vector $\boldsymbol{h}$ are calculated. When analysing models with no control vectors, i.e. with $u_i(t) \equiv 0$ $(i = 1,\ldots,k)$ in 5.1, the generating series expansion reduces to a Taylor series of the output vector function $\boldsymbol{h}$ at the initial time $t = 0$ [12]. By parameterising the unknown initial conditions and therefore regarding them as additional parameters, the Jacobi matrix may also have up to $p+n$ columns. The augmented parameter vector is defined as, $\boldsymbol{\theta} = \binom{\boldsymbol{\theta}}{\boldsymbol{x_0}}$, and the Jacobi matrix calculated by taking the partial derivatives of the constants in the generating series with respect to the unknown parameters in $\boldsymbol{\theta}$ is given by:

$$\frac{\partial \boldsymbol{G}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}) = \begin{pmatrix} \dfrac{\partial \boldsymbol{h}}{\partial \theta_1} & \cdots & \dfrac{\partial \boldsymbol{h}}{\partial \theta_{p+n}} \\ \dfrac{\partial \mathscr{L}_{\boldsymbol{f}_0}\boldsymbol{h}}{\partial \theta_1} & \cdots & \dfrac{\partial \mathscr{L}_{\boldsymbol{f}_0}\boldsymbol{h}}{\partial \theta_{p+n}} \\ \dfrac{\partial \mathscr{L}^2_{\boldsymbol{f}_0}\boldsymbol{h}}{\partial \theta_1} & \cdots & \dfrac{\partial \mathscr{L}^2_{\boldsymbol{f}_0}\boldsymbol{h}}{\partial \theta_{p+n}} \\ \vdots & \ldots & \vdots \end{pmatrix}. \tag{5.10}$$

For models of the form defined in 5.1-5.3, the individual input functions should be incorporated into calculations [9, 20]. An output can be expanded in a Fliess series [19] with respect to time and *inputs*, and the coefficients of this series are $\boldsymbol{h}(\boldsymbol{x}(0),\boldsymbol{\theta})$ and:

$$\mathscr{L}_{\boldsymbol{f}_{j_0}}\ldots\mathscr{L}_{\boldsymbol{f}_{j_q}}\boldsymbol{h}(\boldsymbol{x}(t),\boldsymbol{\theta})|_0, \tag{5.11}$$

where $\boldsymbol{f}_{j_0},\ldots,\boldsymbol{f}_{j_q}$ represent all possible combinations of the vector fields $\{\boldsymbol{f}_j, j = 0,\ldots,k\}$ [9, 12]. The notation $|_0$ indicates that this matrix is evaluated in the point $\boldsymbol{x}(0)$. For example, the Jacobi matrix associated with the full model in 5.1 if $k = 1$, calculated with respect to the unknown parameters in the augmented parameter vector $\boldsymbol{\theta}$ is [12]:

$$\frac{\partial \boldsymbol{G}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial \boldsymbol{h}}{\partial \theta_1} & \cdots & \frac{\partial \boldsymbol{h}}{\partial \theta_{p+n}} \\ \frac{\partial \mathscr{L}_{\boldsymbol{f}_0}\boldsymbol{h}}{\partial \theta_1} & \cdots & \frac{\partial \mathscr{L}_{\boldsymbol{f}_0}\boldsymbol{h}}{\partial \theta_{p+n}} \\ \frac{\partial \mathscr{L}_{\boldsymbol{f}_1}\boldsymbol{h}}{\partial \theta_1} & \cdots & \frac{\partial \mathscr{L}_{\boldsymbol{f}_1}\boldsymbol{h}}{\partial \theta_{p+n}} \\ \frac{\partial \mathscr{L}_{\boldsymbol{f}_1}\mathscr{L}_{\boldsymbol{f}_0}\boldsymbol{h}}{\partial \theta_1} & \cdots & \frac{\partial \mathscr{L}_{\boldsymbol{f}_1}\mathscr{L}_{\boldsymbol{f}_0}\boldsymbol{h}}{\partial \theta_{p+n}} \\ \frac{\partial \mathscr{L}_{\boldsymbol{f}_0}\mathscr{L}_{\boldsymbol{f}_1}\boldsymbol{h}}{\partial \theta_1} & \cdots & \frac{\partial \mathscr{L}_{\boldsymbol{f}_0}\mathscr{L}_{\boldsymbol{f}_1}\boldsymbol{h}}{\partial \theta_{p+n}} \\ \frac{\partial \mathscr{L}_{\boldsymbol{f}_1}\mathscr{L}_{\boldsymbol{f}_1}\mathscr{L}_{\boldsymbol{f}_0}\boldsymbol{h}}{\partial \theta_1} & \cdots & \frac{\partial \mathscr{L}_{\boldsymbol{f}_1}\mathscr{L}_{\boldsymbol{f}_1}\mathscr{L}_{\boldsymbol{f}_0}\boldsymbol{h}}{\partial \theta_{p+n}} \\ \frac{\partial \mathscr{L}_{\boldsymbol{f}_1}\mathscr{L}_{\boldsymbol{f}_0}\mathscr{L}_{\boldsymbol{f}_0}\boldsymbol{h}}{\partial \theta_1} & \cdots & \frac{\partial \mathscr{L}_{\boldsymbol{f}_1}\mathscr{L}_{\boldsymbol{f}_0}\mathscr{L}_{\boldsymbol{f}_0}\boldsymbol{h}}{\partial \theta_{p+n}} \\ \vdots & \cdots & \vdots \\ \frac{\partial \mathscr{L}_{\boldsymbol{f}_{j_0}}\ldots\mathscr{L}_{\boldsymbol{f}_{jq}}\boldsymbol{h}}{\partial \theta_1} & \cdots & \frac{\partial \mathscr{L}_{\boldsymbol{f}_{j_0}}\ldots\mathscr{L}_{\boldsymbol{f}_{jq}}\boldsymbol{h}}{\partial \theta_{p+n}} \\ \vdots & \cdots & \vdots \end{pmatrix}, \tag{5.12}$$

where $j_0, j_1, \ldots, j_q \in [0,1]$. For structural identifiability, it is sufficient for $\frac{\partial \boldsymbol{G}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta})$ to have rank $p+n$, implying that all initial values and system parameters can uniquely be determined. It is known from linear algebra that rank deficiency of a matrix is equivalent to it having a nontrivial null-space [21]. The elements in such a nontrivial null-space reveal the nature of the correlation between the individual unknown parameters. Examples of this will be shown in the following section.

5.3. EXAMPLES

5.3.1. SMALL BENCHMARK MODEL

This small model, first published by Denis-Vidal *et. al.*, shows the role initial conditions play in the the structural identifiability of models with no control input [4]. It comprises 2 state equations:

$$\frac{dx_1}{dt} = \theta_1 x_1^2 + \theta_2 x_1 x_2, \tag{5.13}$$

$$\frac{dx_2}{dt} = \theta_3 x_1^2 + x_1 x_2. \tag{5.14}$$

State x_1 is measured directly and so $y = x_1$. The unknown parameter vector is $\boldsymbol{\theta} = [\theta_1, \theta_2, \theta_3]$. A numerical analysis reveals that the model is structurally unidentifiable

when $x_2(0) = 0$. This is indicated by the distinct gap between the second and third singular values in figure 5.1.

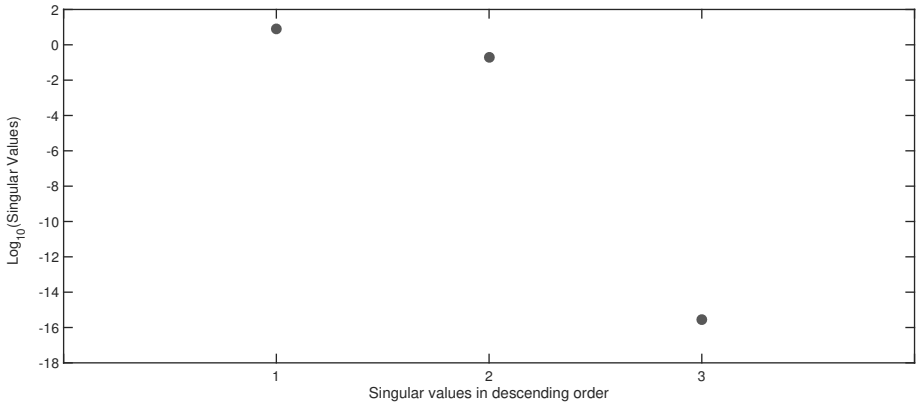


Figure 5.1: **Singular values of the Denis-Vidal model:** When $x_2(0) = 0$ and state x_1 measured, the model is structurally unidentifiable. The distinct gap between the second and third singular values suggests that the sensitivity matrix is rank deficient and that there is 1 set of totally correlated parameters.

Denis-Vidal *et. al.* [4] observed that parameters θ_2 and θ_3 are structurally unidentifiable. This result is confirmed in figure 5.2, where the nonzero entries in the last column of the $\boldsymbol{V}$ matrix, that relates to the single singular value beyond the gap in figure 5.1, indicate that both these parameters are unidentifiable. This result is verified by the symbolically calculated nontrivial null-space, where the Jacobi matrix is computed by only taking partial derivatives with respect to θ_2 and θ_3. The null-space is $\mathcal{N}\left(\frac{d\boldsymbol{G}}{d\boldsymbol{\theta^{unid}}}(\boldsymbol{\theta})\right) = \{-\frac{\theta_2}{\theta_3}, 1\}$, where $\boldsymbol{\theta}^{unid} = \{\theta_2, \theta_3\}$.

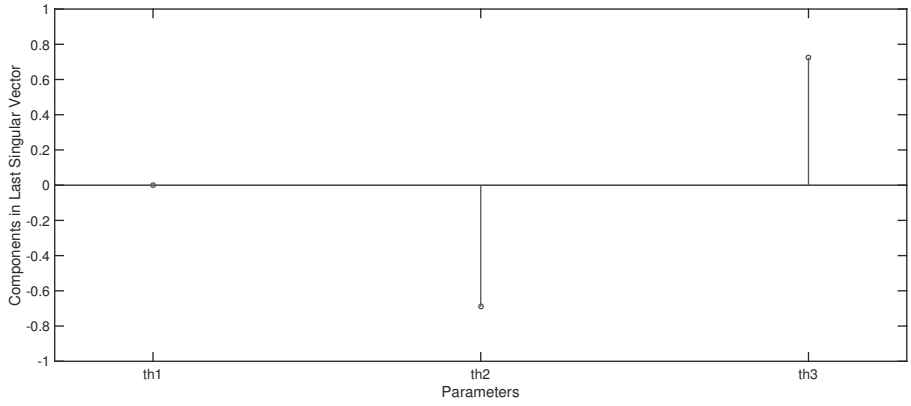


Figure 5.2: **Entries in the last column of the right singular matrix of the Denis-Vidal model:** Nonzero entries indicate that parameters θ_2 and θ_3 are structurally unidentifiable when $y = x_1$ and $x_2(0) = 0$.

5.3.2. BENCHMARK MODEL WITH INPUT

This benchmark model was introduced by Saccomani *et. al.* to investigate the possible connection between a model's structural identifiability and its accessibility [5]. In our analysis, we investigate the structural identifiability of 4 system parameters $\boldsymbol{\theta} = [p_0, p_1, p_2, p_3]$.

$$\frac{dx_1}{dt} = -p_0 u - p_2 x_1 - p_3 x_2, \tag{5.15}$$

$$\frac{dx_2}{dt} = p_3 x_1 x_2 - p_1 x_1. \tag{5.16}$$

State x_1 is measured and so, $y = x_1$. Saccomani *et. al.* showed that when the initial condition $x_2(0) = p_1/p_3$, parameter p_3 becomes structurally unidentifiable. Our numerical results corroborate this result.

For this example, we perform the SVD analysis on the concatenated matrix in 5.17. We change the value of the input u for each sensitivity matrix, $\boldsymbol{S}_i, i = 1, \ldots, k$.

$$\boldsymbol{S}_{all} = \begin{pmatrix} \boldsymbol{S}^1 \\ \boldsymbol{S}^2 \\ \vdots \\ \boldsymbol{S}^k \end{pmatrix}. \tag{5.17}$$

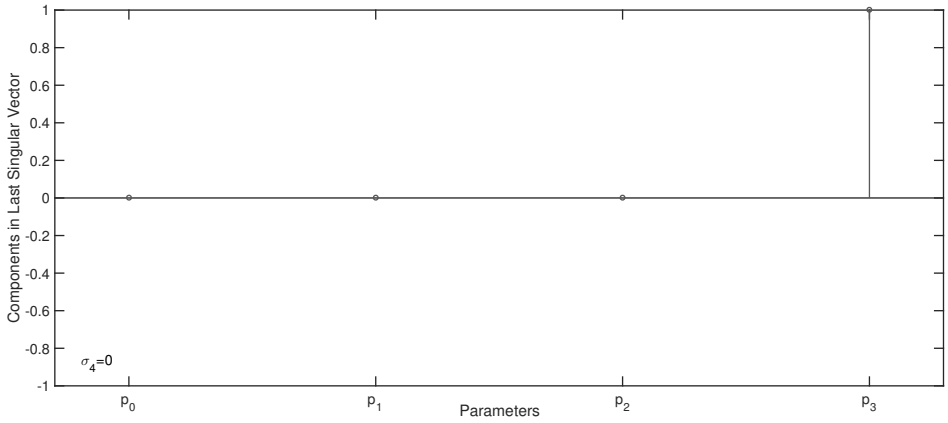


Figure 5.3: **Entries in the last column of the right singular matrix for the benchmark model with input:** When $x_2(0) = p_1/p_3$ and state x_1 measured, parameter p_3 is structurally unidentifiable and the smallest calculated singular value is exactly zero.

Figure 5.3 shows entries of the last vector of the right singular matrix, related to the last singular value of exactly zero. It indicates that parameter p_3 is unidentifiable. This result is verified symbolically. We begin by computing a set of Fliess series coefficient using equation 5.11 as: $\boldsymbol{G}(\boldsymbol{x(0)}, \boldsymbol{\theta}) = \{x_1(0), -p_2 x_1(0) - p_3 x_2(0), -p_0 p_2, -p_0 p_2^2 + p_0 p_3(-p_1 + p_3 x_2(0))\}$. Calculating partial derivatives of the individual elements in this series to the

unknown parameters in $\boldsymbol{\theta}$, and substituting the initial condition, $x_2(0) = p_1/p_3$, the 4×4 Jacobi matrix, where each individual column is related to an unknown parameter is:

$$\frac{d\boldsymbol{G}}{d\boldsymbol{\theta}}(\boldsymbol{\theta}) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & -1 & -1 & 0 \\ -p_2 & 0 & -p_0 & 0 \\ -p_2^2 & 0 & -2p_0p_2 & 0 \end{pmatrix}. \tag{5.18}$$

The fourth column of 5.18 contains only zeros and so parameter p_3 is structurally unidentifiable. Accordingly, the calculated nontrivial null-space of 5.18 is $\mathcal{N}\left(\frac{d\boldsymbol{G}}{d\boldsymbol{\theta}}(\boldsymbol{\theta})\right) = \{0,0,0,1\}$, and this confirms the numerical results in figure 5.3.

5.3.3. MODEL WITH MULTIPLE SETS OF POTENTIAL PROBLEMATIC INITIAL CONDITIONS

A model may have multiple sets of problematic initial conditions and these are not always restricted to zero values. In addition, the choice of parameter values in combination with initial values may also render a model unidentifiable. Consider the following 2 dimensional model, with 1 system parameter and 2 unknown initial conditions, $x_1(0)$ and $x_2(0)$.

$$\frac{dx_1}{dt} = \theta_1 x_1 - x_2, \tag{5.19}$$

$$\frac{dx_2}{dt} = -\theta_1 x_1 + x_2. \tag{5.20}$$

We will see that for certain values of $x_1(0)$ and $x_2(0)$, the model is in steady state, generating no model dynamics and so rendering the model structurally unidentifiable. We analyse this model measuring the output $y = x_1$ and take the unknown parameter vector as $\boldsymbol{\theta} = [\theta_1, x_1(0), x_2(0)]$.

Figure 5.4 shows the smallest singular values of individual sensitivity matrices, each computed on the log scale for different values of $x_1(0)$ and $x_2(0)$ and $\theta_1 = 1$. This figure suggests that the model is structurally unidentifiable when $x_1(0) = x_2(0)$, with shades of blue alluding to singular values of the order 10^{-14} and smaller. The model is unidentifiable even for negative values of $x_1(0)$ and $x_2(0)$. Gap sizes dramatically reduce when $x_1(0) \neq x_2(0)$, indicating that a model's structural identifiablity can easily be reinstated if certain initial values are changed.

Figure 5.5 shows the numerical result obtained at the point [2,2] on the grid and shows that the smallest singular value is numerically equivalent to zero, implying that the sensitivity matrix is rank deficient by 1. The unidentifiable parameters are indicated as nonzero entries in the last column of the right singular matrix, $\boldsymbol{V}$, in figure 5.6.

These results can be verified symbolically by computing the determinant of the 3×3 Jacobi matrix, where each column of this matrix is related to an unknown parameter in $\boldsymbol{\theta}$, θ_1, $x_1(0)$ and $x_2(0)$ respectively. The Jacobi matrix is:

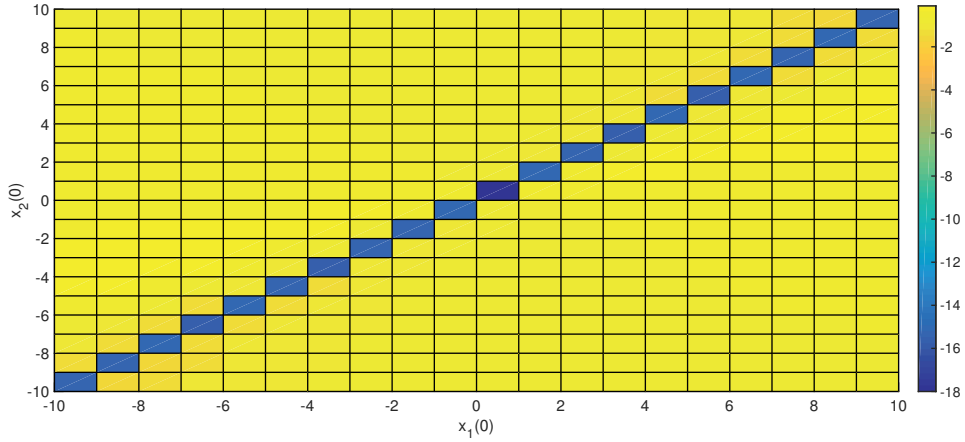


Figure 5.4: **Model with multiple sets of problematic initial conditions:** The small values of the logarithm of the minimum singular values, each calculated as a function of $x_1(0)$ and $x_2(0)$ respectively, suggest that this model has multiple singular points. If we set the initial system parameter value to $\theta_1 = 1$, the model is locally structurally unidentifiable when $x_1(0) = x_2(0)$.

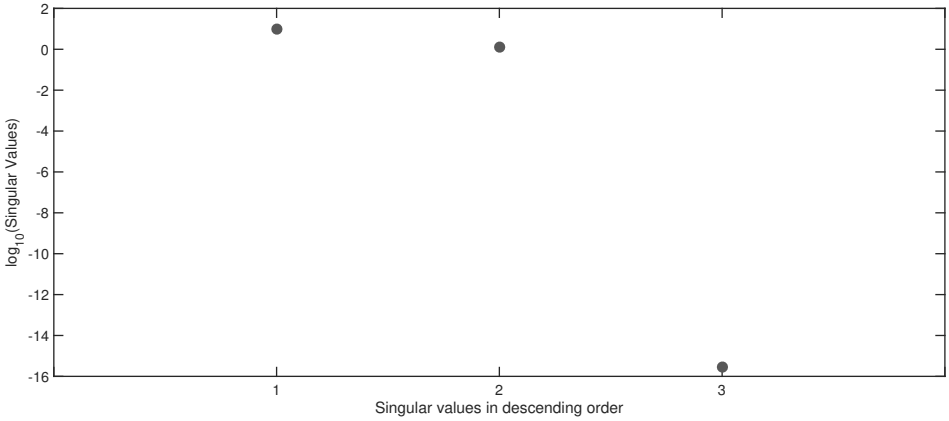


Figure 5.5: **Singular values of the model with multiple sets of problematic initial conditions:** Taking the initial estimate for θ_1 as 1 and $x_1(0) = x_2(0) = 2$, and measuring $y = x_1$, the significant gap in the singular values strongly suggest that the sensitivity matrix is rank deficient.

$$\frac{d\boldsymbol{G}}{d\boldsymbol{\theta}}(\boldsymbol{\theta}) = \begin{pmatrix} 0 & 1 & 0 \\ x_1(0) & \theta_1 & -1 \\ \theta_1 x_1(0) + x_1(0)(1+\theta_1) - x_2(0) & \theta_1(1+\theta_1) & -1-\theta_1 \end{pmatrix}. \tag{5.21}$$

Substituting $\theta_1 = 1$, into 5.21, its determinant is:

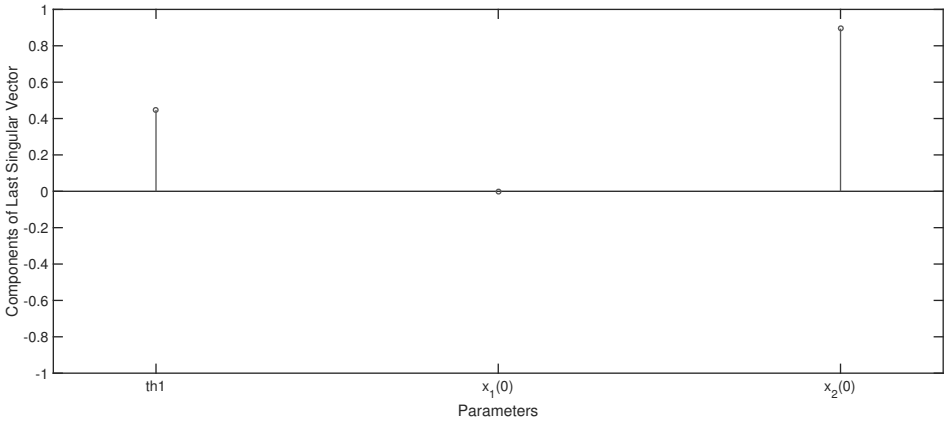


Figure 5.6: **Entries in the last column of the right singular matrix of the model with multiple sets of problematic initial conditions:** Nonzero entries indicate that θ_1 and $x_2(0)$ may be structurally unidentifiable.

$$Det\left[\frac{d\boldsymbol{G}}{d\boldsymbol{\theta}}(\boldsymbol{\theta})\right] = (x_1(0) - x_2(0))^2. \tag{5.22}$$

This indicates that the model will be structurally identifiable when $x_1(0) \neq x_2(0)$. This model, although small and academic, illustrates that models can potentially possess multiple problematic initial conditions and that these may go beyond the obvious $[0,0]$ point.

5.3.4. MODEL DESCRIBING A SIMPLE BIOCHEMICAL NETWORK

In this, the last of the small benchmark models, we consider the 3 state model presented by Villaverde and Banga [3]. The authors show a loss of local structural identifiability for certain initial values. Here, we show that our method is capable of detecting these problematic values and we further consider the notion that problematic values may be associated with steady state conditions. This model describes a simple biochemical network which can be described by the following equations [3, 22]:

$$\frac{dx_1}{dt} = -x_1x_2 + p_2(10 - x_2), \qquad x_1(0) = 0, \tag{5.23}$$

$$\frac{dx_2}{dt} = -x_1x_2 + (p_2 + p_3)(10 - x_2), \qquad x_2(0) = 10, \tag{5.24}$$

$$\frac{dx_3}{dt} = -p_1x_3 + p_3(10 - x_2), \qquad x_3(0) = x_3(0). \tag{5.25}$$

Measuring $\boldsymbol{y} = [x_1, x_3]$ as output, our numerical results reveal that the model might be structurally unidentifiable when analysed using initial conditions $x_1(0) = 0$ and $x_2(0) = 10$. This result was reported by Villaverde and Banga [3]. This possible loss of local structural unidentifiability is indicated in figure 5.7, where there are 2 singular values beyond a large gap. Figure 5.8 depicts the entries in the last 2 columns of the $\boldsymbol{V}$ matrix, related to the 2 singular values. It suggests that parameters p_2 and p_3 are unidentifiable.

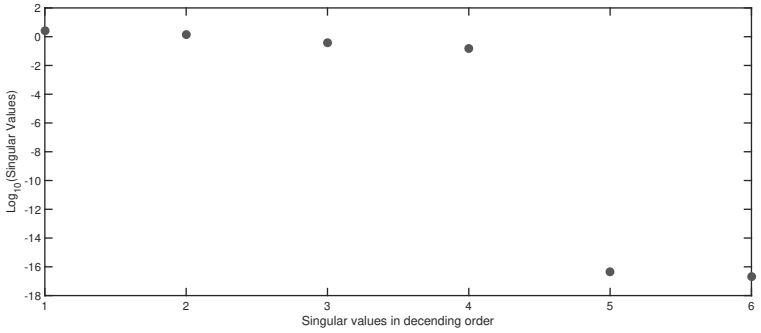


Figure 5.7: **Singular values of the identifiability analysis:** Taking $x_1(0) = 0$ and $x_2(0) = 10$ and measuring the output $\boldsymbol{y} = [x_1, x_3]$, the distinct gap in the singular values suggests that the sensitivity matrix is rank deficient and that there may be structurally unidentifiable parameters.

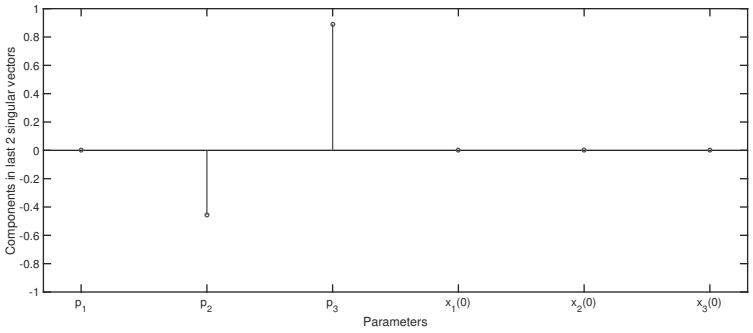


Figure 5.8: **Entries in the last 2 columns of the right singular matrix:** Nonzero entries indicate that parameters p_2 and p_3 are unidentifiable.

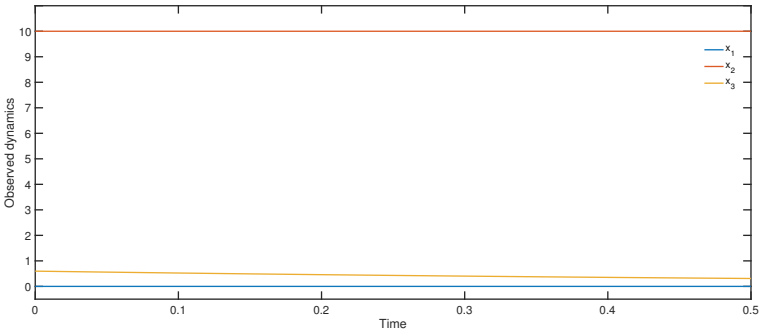


Figure 5.9: **Dynamics observed:** measuring $\boldsymbol{y} = [x_1, x_2, x_3]$ as output, one observes that due to the initial conditions of both $x_1(0)$ and $x_2(0)$, 2 of the observed states remain in steady state. The lack of observed dynamics implies that these initial conditions are contributing to the model's local structural unidentifiability.

This result is confirmed by the nontrivial null-space: $\mathcal{N}\left(\frac{d\boldsymbol{G}}{d\boldsymbol{\theta}^{\boldsymbol{unid}}}(\boldsymbol{\theta})\right) = \{1,0\}\{0,1\}$, where $\boldsymbol{\theta}^{\boldsymbol{unid}} = \{p_2, p_3\}$.

This loss of identifiability can be remedied by choosing different values for either $x_1(0)$ or $x_2(0)$. However, *it cannot* be addressed by measuring state x_2 additionally. Figure 5.9 depicts the observed dynamics when measuring all 3 measurement sensors, $\boldsymbol{y} = [x_1, x_2, x_3]$. It shows that for this special set of initial conditions, no dynamics for either state x_1 or x_2 are observed. This further supports the notion that there is a high probability that models in steady state are structurally unidentifiable and that initial conditions corresponding to a steady states are likely to be problematic.

5.3.5. THREE-PHASE INDUSTRIAL BATCH REACTOR

Here, we consider a model previously used to showcase different reparameterisation and parameter estimation procedures [23, 24]. It comprises the following chemical transformations that are described by 5 reactions, R_{MT}, r_1, r_2, r_3 and r_4 [24]:

$$\begin{aligned} A_s &\rightarrow A_l, & R_{MT}\\ A_l + B &\rightarrow C + D, & r_1\\ B + C &\rightarrow E + D, & r_2\\ B + E &\rightarrow F + D, & r_3\\ B + F &\rightarrow P + D. & r_4 \end{aligned}$$

The first expression describes the dissolution of a solid A, and is governed by the reaction rate R_{MT}. Reactions r_1 to r_4, describe the subsequent steps involved in producing the product P. These reactions are carried out under laboratory conditions and allow for the fast removal of component D. Accordingly, the reverse reactions involving D can be neglected.

These reactions result in the following 7 ODEs [24], with table 5.1 describing the relationships between individual chemical compounds and their mathematical notations:

$$\frac{dx_1}{dt} = -(\theta_0 MW_A)^{1/3}(x_1 MW_A)^{2/3}\frac{\theta_1}{\rho R_{po}}(n_l^{eq} - x_2), \tag{5.26}$$

$$\frac{dx_2}{dt} = (\theta_0 MW_A)^{1/3}(x_1 MW_A)^{2/3}\frac{\theta_1}{\rho R_{po}}(n_l^{eq} - x_2) - \frac{\theta_2 x_3 x_4}{V^2}V, \tag{5.27}$$

$$\frac{dx_3}{dt} = (-\frac{\theta_2 x_3 x_4}{V^2} - \frac{\theta_3 x_3 x_4}{V^2} - \frac{\theta_4 x_3 x_5}{V^2} - \frac{\theta_5 x_3 x_6}{V^2})V, \tag{5.28}$$

$$\frac{dx_4}{dt} = (\frac{\theta_2 x_3 x_4}{V^2} - \frac{\theta_3 x_3 x_4}{V^2})V, \tag{5.29}$$

$$\frac{dx_5}{dt} = (\frac{\theta_3 x_3 x_4}{V^2} - \frac{\theta_4 x_3 x_5}{V^2})V, \tag{5.30}$$

$$\frac{dx_6}{dt} = (\frac{\theta_4 x_3 x_5}{V^2} - \frac{\theta_5 x_3 x_6}{V^2})V, \tag{5.31}$$

$$\frac{dx_7}{dt} = \frac{\theta_5 x_3 x_7}{V^2}V. \tag{5.32}$$

Table 5.1: Molar definition of individual chemical compounds and their mathematical notation.

Symbol	Mathematical substitution
n_{As}	x_1
n_{Al}	x_2
n_B	x_3
n_C	x_4
n_E	x_5
n_F	x_6
n_P	x_7

Parameter θ_0 represents the initial condition of state x_1. The constants MW_A, ρ, R_{po}, n_l^{eq} and V and initial conditions, $x_1(0),\ldots,x_7(0)$, are assumed to be known and so, the identifiability of the 5 remaining parameters, $\boldsymbol{\theta} = [\theta_1,\theta_2,\theta_3,\theta_4,\theta_5]$, is analysed. All initial values, except for $x_1(0)$ and $x_3(0)$, are taken as zero. The measured output defined by Graciano *et. al.* is $\boldsymbol{y} = [x_3, x_7]$ [24].

5

We start by examining the model's directed graph. A directed graph is a graphical representation of an ODE system and depicts the connectivity between individual states. Figure 5.10 shows that the large number of zero-valued initial conditions significantly disrupts the flow of information between the different states. For example, x_5 is completely isolated and if its initial condition was unknown and needed to be estimated, we would have to measure x_5 directly.

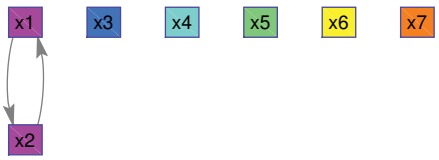


Figure 5.10: **Directed graph of the three-phase industrial batch reactor model:** Assuming that initial conditions $x_2(0) = x_4(0) = x_5(0) = x_6(0) = x_7(0) = 0$, various connections between individual model states, depicted as nodes, are destroyed.

Measuring $\boldsymbol{y} = [x_3, x_7]$, the model is found to be structurally unidentifiable. In this extreme case, all 5 singular values are zero and so the entries in all of the columns of $\boldsymbol{V}$ are considered. The nonzero entries reveal that all parameters are unidentifiable (figure 5.11).

These numerical unidentifiability results were confirmed by the symbolic nontrivial null-space: $\mathcal{N}\left(\frac{d\boldsymbol{G}}{d\boldsymbol{\theta}}(\boldsymbol{x(0)},\boldsymbol{\theta})\right) = \{1,0,0,0,0\}\{0,1,0,0,0\}\{0,0,1,0,0\}\{0,0,0,1,0\}\{0,0,0,0,1\}$, where $\boldsymbol{\theta} = \{\theta_1,\theta_2,\theta_3,\theta_4,\theta_5\}$.

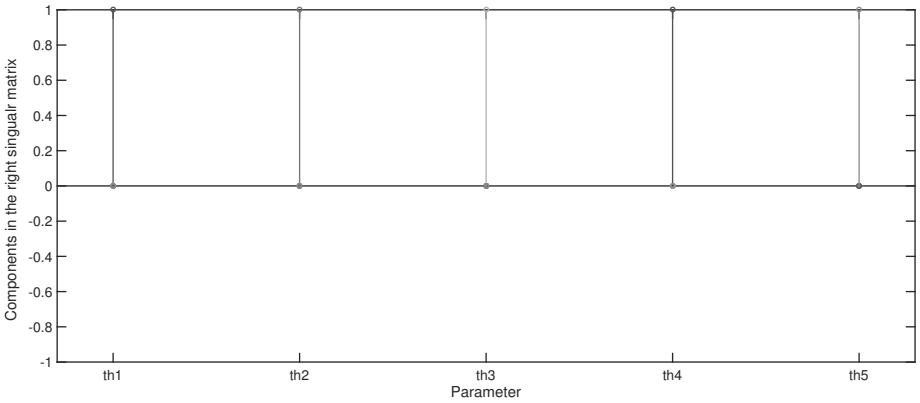


Figure 5.11: **Elements of the right singular matrix, V, of the three-phase industrial batch reactor model:** Considering the entries of all 5 columns of the matrix, each related to a singular value of exactly zero, we see that all unknown parameters are structurally unidentifiable.

To see if the model's structural identifiability can be reinstated by changing certain initial conditions, we now proceed with an iterative analysis of the model. First, we assume that all states are measured to determine whether the model's structural unidentifiability is rooted in the measurement of inadequate states/sensors.

***Analysis 1**: User defined initial conditions measuring all states as output.*

Taking $x_2(0) = x_4(0) = x_5(0) = x_6(0) = x_7(0) = 0$ and measuring the output $\boldsymbol{y} = [x_1, x_2, x_3, x_4, x_5, x_6, x_7]$, the model remains structurally unidentifiable, now with 4 singular values of exactly zero (figure 5.12). The symbolically calculated nontrivial null-space: $\mathcal{N}\left(\frac{d\boldsymbol{G}}{d\boldsymbol{\theta}}(\boldsymbol{x(0)},\boldsymbol{\theta})\right) = \{0,1,0,0,0\}\{0,0,1,0,0\}\{0,0,0,1,0\}\{0,0,0,0,1\}$, where $\boldsymbol{\theta} = \{\theta_1, \theta_2, \theta_3, \theta_4, \theta_5\}$, confirms the numerical results in figures 5.12 and 5.13. This suggests that the source of this model's structural unidentifiability might be its initial conditions and this leads to the second analysis.

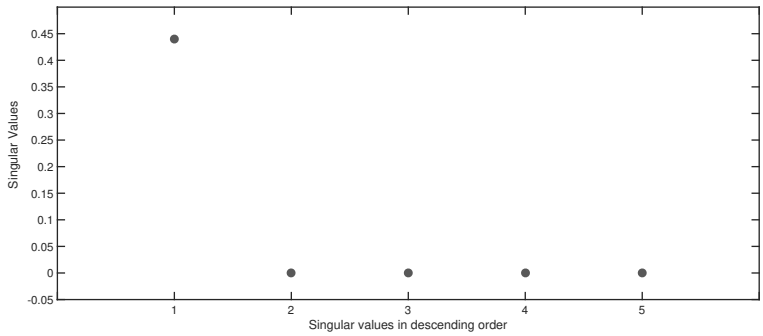


Figure 5.12: **Singular values of the three-phase industrial batch reactor model:** Measuring the output $\boldsymbol{y} = [x_1, x_2, x_3, x_4, x_5, x_6, x_7]$, and assuming initial conditions as $x_2(0) = x_4(0) = x_5(0) = x_6(0) = x_7(0) = 0$, the model remains structurally unidentifiable with 4 singular values beyond the gap.

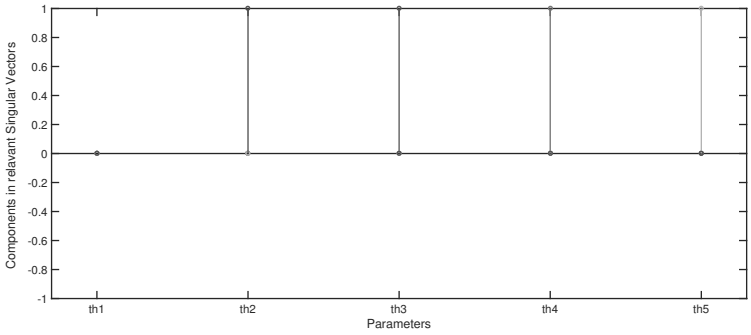


Figure 5.13: **Elements of the last four columns of the right singular matrix, V, for the three-phase industrial batch reactor model:** Each column is related to a singular value beyond the gap in figure 5.12. The nonzero elements suggest that parameters $\theta_2, \theta_3, \theta_4$ and θ_5 are structurally unidentifiable.

***Analysis 2**: All initial conditions are nonzero and measuring $\boldsymbol{y} = [x_3, x_7]$ as output.*
Taking all initial values as nonzero, we find that the model still remains structurally unidentifiable, with 1 singular value of exactly zero. The symbolically calculated non-trivial null-space is: $\mathcal{N}\left(\frac{d\boldsymbol{G}}{d\boldsymbol{\theta}}(\boldsymbol{x(0)}, \boldsymbol{\theta})\right) = \{1,0,0,0,0\}$, where $\boldsymbol{\theta} = \{\theta_1, \theta_2, \theta_3, \theta_4, \theta_5\}$.

These results suggest that to reinstate this model's structural identifiability, one must measure more sensors *and* change initial conditions. For this example, following just one approach is insufficient and in the final analysis, we assume that all states are measured, and proceed by identifying the problematic initial conditions.

***Analysis 3**: Measuring all states and taking more nonzero initial conditions.*
Measuring $\boldsymbol{y} = [x_1, x_2, x_3, x_4, x_5, x_6, x_7]$, the model becomes structurally identifiable when the initial condition of state x_4, along with the original states $x_1(0)$ and $x_3(0)$ are nonzero. To understand this, notice the change in architecture of the model's directed graph in figure 5.14 when $x_4(0) \neq 0$. It results in an increase in the number of connections between the different states.

Table 5.2: **Identifiability results following the extensive analysis of the three-phase industrial batch reactor model**: ✓ indicates experimental conditions under which the model is structural identifiability. × indicates experimental conditions which result in the model's structural unidentifiability, with unidentifiable parameters indicated between brackets.

Analysis	**nonzero Initial Conditions**	**Output, y**	**Structurally Identifiable?**
Original	$x_1(0), x_3(0)$	$[x_3, x_7]$	$\times(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$
Analysis 1	$x_1(0), x_3(0)$	$[x_1, x_2, x_3, x_4, x_5, x_6, x_7]$	$\times(\theta_2, \theta_3, \theta_4, \theta_5)$
Analysis 2	$x_1(0), \ldots, x_7(0)$	$[x_3, x_7]$	$\times(\theta_1)$
Analysis 3	$x_1(0), x_3(0), \boldsymbol{x}_4(0)$	$[x_1, x_2, x_3, x_4, x_5, x_6, x_7]$	✓

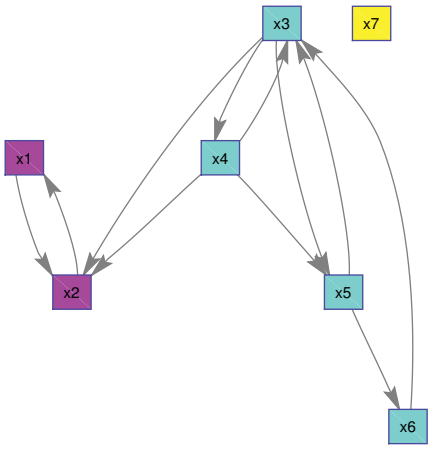


Figure 5.14: **Directed graph of the three-phase industrial batch reactor model with $x_4(0) \neq 0$:** In addition, $x_1(0) = x_3(0) \neq 0$. When the initial condition of state x_4 is no longer zero, various connections between individual model states are reinstated (compare this to figure 5.10). Consequently, there is an increase in the flow of information between the different state equations.

5.3.6. JAK/STAT MODEL

As a final example, we consider the structural properties of the well-known unidentifiable JAK/STAT model. No literature has been published on possible remedies for this. In chapter 3, we presented an efficient method for reparameterising this model. Here, in a bid to offer alternative avenues to obtain its structural identifiability, we identify the origin of its unidentifiability by investigating the model's initial conditions. The constitutive activation of the JAK (Janus kinase)/STAT signalling pathway forms part of both the primary mediastinal B-cell lymphoma (PMBL) and the classical Hodgkin lymphoma (cHL) [25]. Raue *et al.* investigated the identifiability of this benchmark model using three different approaches and concluded that the model is unidentifiable [26]. The initial value of state x_2 is unknown and regarded as an additional parameter and so 23 parameters need to be inferred [26, 27]:

$$\dot{x}_1 = -\theta_1 u_1 c_1 x_1 - \theta_5 x_1 + \theta_6 x_2, \tag{5.33}$$

$$\dot{x}_2 = \theta_5 x_1 - \theta_6 x_2, \tag{5.34}$$

$$\dot{x}_3 = \theta_1 u_1 c_1 x_1 - \theta_2 x_3 x_7, \tag{5.35}$$

$$\dot{x}_4 = \theta_2 x_3 x_7 - \theta_3 x_4, \tag{5.36}$$

$$\dot{x}_5 = \theta_3 x_4 - \theta_4 x_5, \tag{5.37}$$

$$\dot{x}_6 = -\frac{\theta_7 x_3 x_6}{(1+\theta_{13} x_{13})} - \frac{\theta_7 x_4 x_6}{(1+\theta_{13} x_{13})} + \theta_8 c_2 x_7, \tag{5.38}$$

$$\dot{x}_7 = \frac{\theta_7 x_3 x_6}{(1+\theta_{13} x_{13})} + \frac{\theta_7 x_4 x_6}{(1+\theta_{13} x_{13})} - \theta_8 c_2 x_7, \tag{5.39}$$

$$\dot{x}_8 = -\theta_9 x_8 x_7 + c_2 \theta_{10} x_9, \tag{5.40}$$
$$\dot{x}_9 = \theta_9 x_8 x_7 - c_2 \theta_{10} x_9, \tag{5.41}$$
$$\dot{x}_{10} = \theta_{11} x_9, \tag{5.42}$$
$$\dot{x}_{11} = -\theta_{12} c_1 u_1 x_{11}, \tag{5.43}$$
$$\dot{x}_{12} = \theta_{12} c_1 u_1 x_{11}, \tag{5.44}$$
$$\dot{x}_{13} = \frac{\theta_{14} x_{10}}{(\theta_{15} + x_{10})} - \theta_{16} x_{13}, \tag{5.45}$$
$$\dot{x}_{14} = \theta_{17} x_9. \tag{5.46}$$

The model output contains 5 additional unknown parameters:

$$y_1 = x_1 + x_3 + x_4, \tag{5.47}$$
$$y_2 = \theta_{18}(x_3 + x_4 + x_5 + x_{12}), \tag{5.48}$$
$$y_3 = \theta_{19}(x_4 + x_5), \tag{5.49}$$
$$y_4 = \theta_{20} x_7, \tag{5.50}$$
$$y_5 = \theta_{21} x_{10}, \tag{5.51}$$
$$y_6 = \theta_{22} x_{14}, \tag{5.52}$$
$$y_7 = x_{13}, \tag{5.53}$$
$$y_8 = x_9. \tag{5.54}$$

The initial values of the individual model states are $\boldsymbol{x}(0) = \{1.3, \theta_{23}, 0, 0, 0, 2.8, 0, 165, 0, 0, 0.34, 0, 0, 0\}$ [26]. The constants c_1, c_2 and model input u_1 are known and regarded as constant. Let us start by considering the directed graph of the model structure related to this particular set of initial conditions. If one takes a nonzero value for the unknown initial condition $x_2(0)$, the model structure is shown in figure 5.15.

The SVD of the sensitivity matrix reveals that the model is indeed structurally unidentifiable. This is evident from the large gap between the singular values seen in figure 5.16.

The 2 singular values beyond the gap suggest that the null-space contains 2 base vectors and so there are 2 sets of totally correlated parameters. The union of the elements in these 2 sets, $\boldsymbol{\theta}^{unid} = \{\theta_{11}, \theta_{15}, \theta_{17}, \theta_{21}, \theta_{22}\}$, follows from the nonzero elements in figure 5.17. The symbolically calculated nontrivial null-space confirms the results in figures 5.16 and 5.17. Analysing the potential unidentifiability of these parameters symbolically, we calculate the 2 base vectors spanning the null-space as: $\mathcal{N}\left(\frac{d\boldsymbol{G}}{d\boldsymbol{\theta}^{unid}}(\boldsymbol{\theta})\right) = \{0, 0, -\theta_{17}/\theta_{22}, 0, 1\}\ \{-\theta_{11}/\theta_{21}, -\theta_{15}/\theta_{21}, 0, 1, 0\}$, where $\boldsymbol{\theta}_0^{unid} = \{\theta_{11}, \theta_{15}, \theta_{17}, \theta_{21}, \theta_{22}\}$.

We now proceed with an analysis of this model to establish which initial experimental conditions result in the model's structural unidentifiability. If changing the identified problematic initial conditions is experimentally viable, the model's structural identifiability can easily be re-established. If not, the modeller can consider the alternative options listed in chapter 3.

***Analysis 1**: Measuring all states as output.*
Adding additional states to the measured output therefore measuring, $\boldsymbol{y} = [x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_1 + x_3 + x_4, \theta_{18}(x_3 + x_4 + x_5 + x_{12}), \theta_{19}(x_4 + x_5), \theta_{20} x_7,$

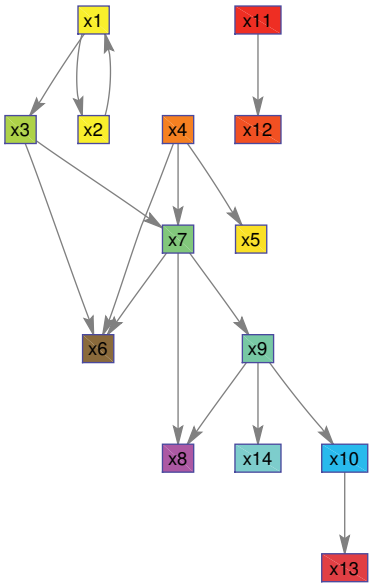


Figure 5.15: **Directed graph of the JAK/STAT model:** This graph is generated for the initial conditions: $x_3(0) = x_4(0) = x_5(0) = x_7(0) = x_9(0) = x_{10} = x_{12}(0) = x_{13}(0) = x_{14}(0) = 0$.

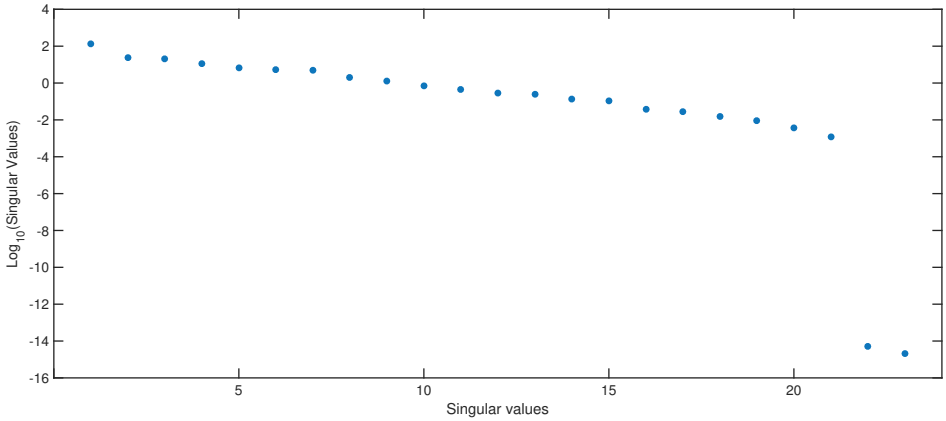


Figure 5.16: **Singular values of the JAK/STAT model:** The 2 singular values beyond the gap suggest that the model is structurally unidentifiable. These values also suggest that there are 2 sets of totally correlated parameters.

$\theta_{21}x_{10}, \theta_{22}x_{14}]$, the model turns out to be structurally identifiable. This result indicates that this model's lack of identifiability could be addressed by measuring additional states/ sensors. These states can be determined using our minimal output set algorithm presented in chapter 2. It is important to keep in mind that measuring additional states/

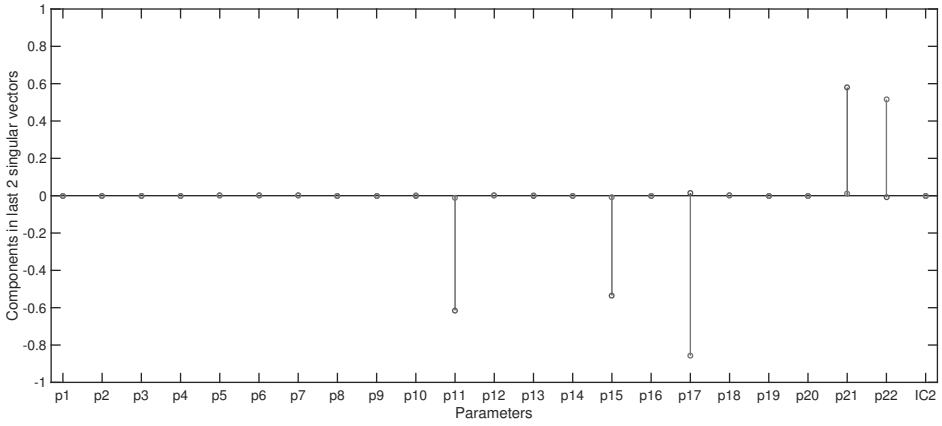


Figure 5.17: **Last 2 columns of the right singular matrix of the JAK/STAT model:** These columns are related to the 2 singular values beyond the gap in figure 5.16. The nonzero elements indicate that parameters $\theta_{11}, \theta_{15}, \theta_{17}, \theta_{21}$ and θ_{22} might be structurally unidentifiable.

5

sensors may have practical experimental limitations.

Let us now investigate the potential role of the model's initial conditions in its unidentifiability.

Analysis 2*: Nonzero initial conditions measuring as output,* $\boldsymbol{y} = [x_1 + x_3 + x_4, \theta_{18}(x_3 + x_4 + x_5 + x_{12}), \theta_{19}(x_4 + x_5), \theta_{20}x_7, \theta_{21}x_{10}, \theta_{22}x_{14}, x_{13}, x_9]$.
When all initial conditions are nonzero, the model also turns out to be structurally identifiable even when measuring the output defined in Raue *et. al.* [26]. This shows that the model's unidentifiability can also be attributed to its initial conditions and urges us to look for the initial conditions that cause the identifiablity problem.

Analysis 3*: Determining which zero initial conditions to avoid.*
By performing an exhaustive search and selecting different combinations of nonzero initial values, we are able to identify individual or groups of problematic initial conditions. We find that if $x_{10}(0)$ *and* $x_{14}(0)$, along with the original states $x_1(0), x_2(0), x_6(0)$ and $x_8(0)$ are nonzero, the model turns out to be structurally identifiable when measuring the defined output $\boldsymbol{y} = [x_9, x_{13}, x_1 + x_3 + x_4, \theta_{18}(x_3 + x_4 + x_5 + x_{12}), \theta_{19}(x_4 + x_5), \theta_{20}x_7, \theta_{21}x_{10}, \theta_{22}x_{14}]$.

For some smaller models, one may identify certain problematic initial conditions by looking at a model's state equations and/or directed graph (see figures 5.10 and 5.14 for example). However, the problematic initial conditions of large models cannot be identified by merely looking at the model's state equations and/or directed graph. For this model, there is no change in the structure of the directed graph when $x_{10}(0) = x_{14}(0) \neq 0$. One therefore requires a method capable of detecting these initial conditions for large models in an computationally efficient way. The results of these analyses are summarised in table 5.3.

Table 5.3: **Identifiability results following the extensive analysis of the JAK/STAT model**: ✓ indicates initial conditions for which the model is structural identifiability. × indicates conditions which result in structural unidentifiability, with unidentifiable parameters indicated between brackets.

Analysis	**Nonzero Initial Conditions**	**Output, y**	**Identifiable?**
Original	$x_1(0), x_2(0), x_6(0), x_8(0), x_{11}(0)$	$[x_1+x_3+x_4, \theta_{18}(x_3+x_4+x_5+x_{12}), \theta_{19}(x_4+x_5), \theta_{20}x_7, \theta_{21}x_{10}, \theta_{22}x_{14}, x_{13}, x_9]$	$\times(\theta_{11}, \theta_{15}, \theta_{17}, \theta_{21}, \theta_{22})$
Analysis 1	$x_1(0), x_2(0), x_6(0), x_8(0), x_{11}(0)$	$[x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_1+x_3+x_4, \theta_{18}(x_3+x_4+x_5+x_{12}), \theta_{19}(x_4+x_5), \theta_{20}x_7, \theta_{21}x_{10}, \theta_{22}x_{14}, x_{13}, x_9]$	✓
Analysis 2	$x_1(0), \ldots, x_{14}(0)$	$[x_1+x_3+x_4, \theta_{18}(x_3+x_4+x_5+x_{12}), \theta_{19}(x_4+x_5), \theta_{20}x_7, \theta_{21}x_{10}, \theta_{22}x_{14}, x_{13}, x_9]$	✓
Analysis 3	$x_1(0), x_2(0), x_6(0), x_8(0), \mathbf{x_{10}(0)}, x_{11}(0), \mathbf{x_{14}(0)}$	$[x_1+x_3+x_4, \theta_{18}(x_3+x_4+x_5+x_{12}), \theta_{19}(x_4+x_5), \theta_{20}x_7, \theta_{21}x_{10}, \theta_{22}x_{14}, x_{13}, x_9]$	✓

5.4. CONCLUSIONS

This chapter offers a modeller the possibility to make his/her model structurally identifiable simply by choosing alternative initial conditions. In the last example, we showed that problematic initial conditions cannot merely be found by regarding a model's directed graph and state equations. This is particularly true for large ODE models and so, one requires an efficient method capable of detecting these problematic values.

We showed that our sensitivity based method, which numerically assesses the local structural identifiability of a model, correctly and efficiently identifies problematic initial conditions. Our method can do this analysis within seconds, and in doing so, we go one step further than the general identifiability methods, in the sense that we are capable of detecting the cause of a model's unidentifiability. This is useful to experimental modellers wishing to address the structural unidentifiability of their models.

The key findings of this chapter can be summarised as: 1) a model may have multiple sets of problematic initial conditions, 2) problematic initial conditions are not restricted to zero values, and 3) certain initial conditions may result in steady state conditions. Moreover, we saw that there is a high probability that models in steady state are structurally unidentifiable and so initial conditions corresponding to a steady state are likely

to be problematic.

REFERENCES

[1] K. Z. Yao, B. M. Shaw, B. Kou, K. B. McAuley, and D. W. Bacon, *Modeling ethylene/butene copolymerization with multi-site catalysts: Parameter estimability and experimental design,* Polymer Reaction Engineering **11**, 563–588 (2003).

[2] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmuller, and J. Timmer, *Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood,* Bioinformatics **25**, 1923–1929 (2009).

[3] A. Villaverde and J. Banga, *Structural properties of dynamic systems biology models: Identifiability,reachability and initial conditions,* Processes **5** (2017), 10.3390/pr5-20029.

[4] L. Denis-Vidal, G. Joly-Blanchard, and C. Noiret, *Some effective approaches to check the identifiability of uncontrolled nonlinear systems,* Math. Comput. Simul. **57**, 35 (2001).

[5] M. P. Saccomani, S. Audoly, and L. D'Angió, *Parameter identifiability of nonlinear systems: the role of initial conditions,* Automatica **39**, 619 (2003).

[6] J. Karlsson, M. Anguelova, and M. Jirstrand, *An efficient method for structural identifiability analysis of large dynamic systems,* IFAC Proceedings Volumes **45**, 941 (2012), 16th IFAC Symposium on System Identification.

[7] G. Bellu, M. P. Saccomani, S. Audoly, and L. D'Angió, *Daisy: A new software tool to test global identifiability of biological and physiological systems,* Computer Methods and Programs in Biomedicine **81**, 52 (2007).

[8] A. F. Villaverde, A. Barreiro, and A. Papachristodoulou, *Structural identifiability of dynamic systems biology models,* PLOS Computational Biology **20**, 1 (2016).

[9] E. Walter and L. Pronzato, *On the identifiability and distinguishability of nonlinear parametric models,* Mathematics and Computers in Simulation **42**, 125 (1996).

[10] M. Henson and D. Seborg, *Nonlinear Process Control* (New Jersey: Prentice Hall, 1997).

[11] J. D. Stigter and R. L. M. Peeters, *On a geometric approach to the structural identifiability problem and its application in a water quality case study,* in *2007 European Control Conference (ECC)* (2007) pp. 3450–3456.

[12] J. D. Stigter and J. Molenaar, *A fast algorithm to assess local structural identifiability,* Automatica **58**, 118 (2015).

[13] J. D. Stigter, D. Joubert, and J. Molenaar, *Observability of complex systems: Finding the gap,* Scientific Reports **7**, 1 (2017).

[14] Y. Bard, *Nonlinear Parameter Estimation* (Academic Press Inc, 1974).

[15] M. Vidyasagar, *Nonlinear systems analysis* (Prentice Hall, Englewood Cliffs, NJ, 1993).

[16] A. Gábor, A. F. Villaverde, and J. R. Banga, *Parameter identifiability analysis and visualization in large-scale kinetic models of biosystems,* BMC Systems Biology **11**, 1 (2017).

[17] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th ed. (The Johns Hopkins University Press, 2013).

[18] G. Quintana-Ortí and E. S. Quintana-Ortí, *Parallel codes for computing the numerical rank,* Linear Algebra and its Applications **275-276**, 451 (1998).

[19] E. T. Tunali and T. J. Tarn, *New results for identifiability of nonlinear systems,* IEEE Transactions on Automatic Control **32**, 146 (1987).

[20] A. Iggidr, *Controllability, Observability, and Stability of Mathematical Models,* in *Encyclopedia of Life Support Systems (EOLSS)*, Vol. Mathematical Models, edited by J. A. Filar (UNESCO, Eolss Publishers, 2004).

[21] J. D. Stigter, M. B. Beck, and J. Molenaar, *Assessing local structural identifiability for environmental models,* Environmental Modelling and Software **93**, 398 (2017).

[22] E. August and A. Papachristodoulou, *A new computational tool for establishing model parameter identifiability,* Journal of Computational Biology **6**, 875 (2009).

[23] A. Ben-zvi, *Reparameterization of inestimable systems with applications to chemical and biochemical reactor systems,* AIChE Journal **54**, 1270–1281 (2008).

[24] J. Graciano, D. Mendoza, and G. L. Roux, *Performance comparison of parameter estimation techniques for unidentifiable models,* Computers and Chemical Engineering **64**, 24–40 (2014).

[25] V. Raia, M. Schilling, M. Böhm, B. Hahn, A. Kowarsch, A. Raue, C. Sticht, S. Bohl, M. Saile, P. Möller, N. Gretz, J. Timmer, F. Theis, W.-D. Lehmann, P. Lichter, and U. Klingmüller, *Dynamic mathematical modeling of IL13-induced signaling in hodgkin and primary mediastinal B-cell lymphoma allows prediction of therapeutic targets,* Cancer Research **71**, 693 (2011).

[26] A. Raue, J. Karlsson, M. P. Saccomani, M. Jirstrand, and J. Timmer, *Comparison of approaches for parameter identifiability analysis of biological systems,* Bioinformatics **30**, 1440–1448 (2014).

[27] D. V. Raman, *On the Identifiability of Highly Parameterised Models of Physical Processes*, Ph.D. thesis, University of Oxford (2016).

6

General Discussion

Dominique Joubert

Some of the big challenges that we face, both societal and scientific, are just not solvable by people sitting in their single-discipline silos – bringing those disciplines together in the long term is what provides the big, big breakthroughs.

(Kedar Pandya, 2011)

6.1. INTRODUCTION

In this thesis, I endeavoured to address the following quote...

"The lack of real contact between mathematics and biology is either a tragedy, a scandal or a challenge, it is hard to decide which." (Gian-Carlo Rota, 1932-1999)

It was briefly mentioned in chapter 1 that advances in the field of systems biology may be stifled by the lack of development in technologies that are required to analyse models from this field [1]. The job at hand is clear, scientists require easy to implement methods to help them strengthen their scientific outcomes. Ideally, these methods should ease the computational burden often associated with reaching these outcomes. The solution lies in an interdisciplinary approach, and the work presented in this thesis bridges this gap by drawing from both mathematics and control engineering.

Before using a model, its identifiability should first be analysed. This analysis can be divided into two categories, structural and practical, with the former inspected before any experiments are conducted. Structural identifiability is important since predictions regarding states and inputs of unidentifiable models may be incorrect. These significant consequences are even more poignant in the context of systems biology, where inaccurate predictions involving medical and pharmaceutical models, can lead to incorrect diagnoses and treatment regimens [2].

Structural identifiability and the need for efficient methods to analyse it is the main theme of this thesis and the silver thread connecting the individual chapters. I systematically showed how researchers can use the hybrid structural identifiability method, developed by Stigter and Molenaar [3], in an array of *complimentary applications*. These range from the preliminary design of experiments, to practical examples of how one is to go about eliminating redundant parameters from a model. I chose to use this method since it *efficiently* assesses a model's identifiability by first analysing it *numerically*, and then uses the numerical results in subsequent *symbolic calculations*. This significantly reduces the computational demand often associated with symbolic analyses. In this chapter, I highlight my contributions to both the systems biology and structural identifiability communities.

THE MODEL DEVELOPMENT PROCESS

The role and location of a structural identifiability analysis in the model development process is shown in figure 1.1, with an alternative structure given in [4]. Ideally, this analysis should always be performed since its conclusions have significant repercussions on the rest of the virtuous cycle. In the worst case scenario, a practical identifiability analysis may reveal that certain parameters cannot be estimated uniquely. Moreover, some parameters may be totally correlated. The source of this problem may however not be understood. Consequently, a modeller can be left with very little to show after spending significant time, effort and money on the development of his/her model. Despite its importance, numerous misconceptions and challenges regarding structural identifiability remain and these include:

1. *Ignoring structural identifiability analysis.*
Systems biology models are occasionally published without any mention of a structural identifiability analysis. For these models, practical identifiability results might reveal that the degree of correlation between individual parameters is admissible and so, a model can be adopted. However, if this is not the case, the discovery of totally correlated parameters within a model then only comes after incurring costs related to experimental measurements.

   An example of such a model is a recently published respiratory mechanics model [5]. In their paper, the authors conclude that the model is indeed practically unidentifiable. They suggest that a possible remedy to this is to include a procedure which limits a patient's air supply and so, additional experiments are required. A structural identifiability analysis of this model might be insightful and avoid the fact that in the future, additional experiments may again be needed.

   A structural identifiability analysis was also omitted in the development of a biodiesel model by Price *et. al.* [6] and the subsequent paper written on the topic of experimental design by Yu *et. al.* [7]. In both of these papers, the practical identifiability of the model is analysed and the authors conclude that only 10 of the 20 system parameters can be estimated accurately. However, a structural identifiability analysis would reveal that when the initial condition of x_{15} [7] is zero, the model is structurally unidentifiable. Crucially, the experimental initial value of this chemical constituent is 0.0000097165 and given the model's sensitivity to small values of this particular state's initial condition, this small value is most probably the cause of this model's practical identifiability issues. Given this insight, experimental design strategies might initially be focused only on changing the initial value of the chemical compound related to x_{15}. This step might be sufficient in addressing the model's practical identifiability issues.

2. *Structurally unidentifiable, now what?*
Papers that do cover the topic of structural identifiability, and so acknowledge its importance in the model development process, often finish with an unidentifiable conclusion. With this abrupt end to the story, these papers all play a vital yet limited role in the model development process and the issue of solving the "problem" still remains. Conversely, examples of publications that mention possible strategies to regain a model's structural identifiability include: Anguelova *et. al.* [8] (determining minimal output sets), Saccomani *et. al.* [9] (determining the role of initial conditions in a model's local structural identifiability), Villaverde *et. al.* [10]. In the work done by Chappell and Gunn [11] and Evans and Chappell [12], the authors identify identifiable parameter combinations and reparameterise unidentifiable models.

3. *A model is too large or contains irrational functions.*
The number of software applications capable of analysing large nonlinear ODE models is limited. Villaverde *et. al.* commented on this by stating that a structural identifiability analysis is seldom performed since the computational demand prohibits the analysis of large models [10]. The methods currently available for the analysis of large systems are: 1) a semi-numerical Exact Arithmetic Rank approach

proposed in 2012 by Karlsson *et. al.* [13] 2) STRIKE-GOLDD [10] 3) Profile likelihood approach [14] and 4) the hybrid method by Stigter *et. al.*, where large models were analysed in [15]. Methods (1)-(3) are computationally demanding, and so the method in (4) may be preferable when analysing certain large models.

Methods are often also restricted to the analysis of rational models. An example of a model not analysed for its structural properties due to the fact that it contains irrational functions is the gastrointestinal stromal tumor (GIST) metastasis to the liver model [16]. In their publication, the authors explicitly mention the fact that they could not analyse the model's structural identifiability since most methods only facilitate the analysis of rational models. For these cases, the method in (4) may be of use.

4. *Simplicity and availability of methods.*
Finally, a possible reason for failing to analyse a model's structural identifiability might be the limited number of readily available software applications. The available open-source software toolboxes include COMBOS [17], DAISY [18], EAR [13], STRIKE-GOLDD [10] and GenSSI 2.0 [19]. According to Ligon *et. al.* [19], one of the shortcomings associated with some of these methods is the fact that they do not support community standards, i.e. the Systems Biology Markup Language (SBML), and only support the analysis of individual experimental conditions. A method should therefore be capable of analysing ODE models in both SBML and general formats. Some of the above mentioned methods also only allow for the analysis of *small* rational ODE models and this further contributes to the fact that some of these cannot be used by the systems biology community.

6.2. HIGHLIGHTS OF RESULTS

The highlights of the results obtained in this thesis include:

1. *Ignoring structural identifiability analysis - A missed opportunity.*
Structural identifiability is a prerequisite for practical identifiability and so it is always good modelling practice to conduct this analysis. The structural analysis of a model can ensure that no experimental effort is lost by ensuring that all parameters can be estimated. By failing to conduct such an analysis, one misses out on the opportunity of preliminary designing experiments. More concisely, in chapters 2 and 5 we showed that our method can be used to determine which states/sensors should be measured and which experimental initial conditions should be avoided.

2. *Structurally unidentifiable, here's what?*
Offering modellers insight into the source of their model's unidentifiability and indicating how they are to go about reinstating it is a prominent theme in this thesis. In chapter 3 for example, we addressed the topic of reparameterising structurally unidentifiable models and in chapter 5 we showed that our method is capable of detecting problematic initial conditions for large ODE models.

3. *Large models and/or irrational functions.*
The hybrid nature of our method allows for the analysis of large ODE models. Ex-

amples of these can be found in chapter 4 where we analysed, amongst others, the large Chinese Hamster model. Moreover, in chapter 2 we determined the MOS of an irrational model (example 8).

4. *A simple and accessible method - preparation.*
The final important contribution of this thesis is the work done in a bid to understand the main factors that influence our numerical results. The insights gained in chapter 4 will, in the future, help us develop an easy-to-use open-source software application that will be able to analyse a plethora of different models from various disciplines. It will include, but not be limited to, the SBML markup language.

6.3. DISCUSSION

MINIMAL OUTPUT SETS

This application of our method addressed the first objective listed in chapter 1, identifying the minimal sets of outputs that need to be measured to ensure a model's structural identifiability. Minimal output sets form an integral part of our structural identifiability method. As presented in chapter 2, it can be used in the optimal design of experimental measurement sets. This ensures that a scientist knows exactly which sensors he/she should measure to ensure that the model in question remains structurally identifiable. This can indeed be a valuable insight especially when numerous of these sets exist, in which case he/she can take physical constraints such as experimental cost and time into account.

A second important application of the minimal output set algorithm is in the structural identifiability method itself. This was alluded to in chapter 4, where potential problems that may be encountered when using our method were mentioned (section 4.3.3). We showed that when a model is analysed measuring only a small number of states/sensors in proportion to the total number of states of the model, the numerical accuracy of our method may be compromised. In these instances a model should be analysed thoroughly, with the explicit aim of determining both which states/sensors are important and whether or not they are present in the defined measured output. To this end, our MOS algorithm developed in chapter 2 should be used.

Even when a model is found to be structurally unidentifiable, performing a MOS analysis can be valuable in identifying the potential cause of this unidentifiability. A good understanding of a model's minimal output sets can help a scientist make a well-informed decision on how to reinstate the model's identifiability by providing a list of possible states/sensors that should be measured.

The final important point to raise in this concluding discussion is the matter of improving the efficiency of the MOS algorithm itself. The aim is to reduce the number of analyses that are required during an exhaustive search for a model's minimal output sets. In the supplementary file S9 in the Appendix of chapter 2, we introduced the concept of randomly omitting states/sensors from an output. We subsequently showed that by repeating this process, one can significantly reduce the number of analyses required to detect all the states/sensors that should be included into a measured set, whilst maintaining a 99.5% probability of detecting such an important set. The equation used to

compute the number of Bernoulli trials required, R, is:

$$\bar{P}_{det} = (1 - P(X = K))^R, \tag{6.1}$$

where $\bar{P}_{det}$ is the probability of *not* detecting a set of states/sensors that should be included into a model's minimal sensor set. This value should be small and we used 0.5% as a benchmark. $P(X = K)$ indicates the probability of successfully detecting a set that consists of K important states/sensors that, if not measured, result in unidentifiability. R is the required number of Bernoulli trials. The novelty of the work presented in this chapter is that one could determine the sets of important sensors, even for large systems biology models within a matter of minutes.

REPARAMETERISING UNIDENTIFIABLE MODELS

This addressed the second objective listed in chapter 1. In chapter 3 we set out to present a method that could provide theoretical suggestions for the reparameterisation of structurally unidentifiable models. The limited number of detailed publications on this topic have, thus far, only included small models comprising 2 or 3 state equations as examples. These include the work of authors such as Neil Evans and Mike Chappell and Nicolette Meshkat and Marisa Eisenberg, [11, 12, 17, 20].

We obtained reparameterisations for large ODE models due to two prominent features of our hybrid method. The first is that unknown initial conditions can be treated on the same footing as parameters. Accordingly, repameterisations involving unidentifiable initial conditions could easily be obtained. The second is the fact that the numerical analysis of a model filters out suspected unidentifiable parameters. This significantly reduces the number of subsequent symbolic calculations that are required to obtain suitable solutions to the partial differential equations describing the linear dependence between the columns of the Jacobi matrix. For example, we reparameterised the lung cancer model with 21 states and 75 system parameters in hardly any time. This reparameterisation necessitated the transformation of certain states.

THE ROLE OF INITIAL CONDITIONS

In chapter 5 we addressed the third thesis objective, i.e. detecting sets of problematic initial conditions. We showed that our algorithm can be added to a list of 3 software packages capable of detecting these problematic values and that it can do so for large ODE models. This is novel since previous publications on this topic have only included small 2 state models as examples.

The numerical characteristics of our method, which requires the definition of both system and initial values, enabled us to search for potential sets of initial conditions leading to a model's structural unidentifiability. This was done by performing an exhaustive search and ultimately, a model's identifiability could be reinstated by changing the values of the identified culprits.

The ability to efficiently detect problematic initial conditions could also complement the preliminary experimental design phase, by allowing for the compilation of a list of initial conditions that should ideally be avoided. As we saw in chapter 5, these values can be zero or nonzero or combinations of initial and parameters values, and that a model may have multiple sets of these problematic values. An important point that became

clear in this discussion was that models in steady state have a great chance to be structurally unidentifiable.

A ROBUST NUMERICAL METHOD

Ultimately, the aim is to achieve the above mentioned functionally for *a wide range* of ODE models. Our method should therefore be robust. In chapter 4 we addressed the final remaining objective defined in chapter 1, identify key factors that influence the numerical structural identifiablity results. We performed a sensitivity analysis of our sensitivity based method to determine which factors where the most influential on the numerical accuracy.

We identified: a. parameter and initial values and, b. the composition of the measured output vector as 2 factors that had a significant effect on results. When analysing a model under specific experimental conditions, both these factors can be fixed. This implies that a researcher may wish to analyse his/her model for a specific set of initial conditions, and/or a specific measured output. We therefore required a strategy to mitigate this sensitivity.

To this end, we introduced the concept of vertically concatenating numerous sensitivity matrices. Each of these matrices are evaluated for a different set of parameter and initial values, keeping key user defined values and the values of possible unidentifiable parameters unchanged. Here the MOS algorithm plays a crucial role, since it can be used to first detect possible unidentifiable parameter sets. Accordingly, the MOS algorithm can significantly contribute to the robustness of our identifiability method.

6.4. CONCLUSION AND FUTURE WORK

To conclude, in this thesis I introduced an algorithm that can be used in an array of useful applications during the model development process. These include: 1) determining minimal output sets, 2) reparameterising structurally unidentifiable models and 3) detecting problematic initial conditions. Each of these can be implemented *before any experiments are conducted* and can play a potential role in the optimisation of the modelling process.

Future developments in the field of systems biology will continue to drive and define the direction of development of the co-disciplinary research fields that support it. In the structural identifiability context, the need for numerous advances remains, the first of which is the development of a *database* that contains models of different sizes. This can in turn be referred to when testing and developing new structural identifiability methods. Large amounts of time can often be spent on finding relevant ODE models with well defined initial conditions and measured outputs and this small step might help. It may also contribute to the standardisation of results, with newly published methods including some benchmark examples that allow for comparison with pre-existing methods.

A second major contribution that remains to be made is that of easy-to-use open-source software. The availability of such tools will allow model developers to analyse their respective models under different experimental conditions as well as offer them the opportunity of implementing the 3 different applications presented in this thesis. The hope is that making this software readily available will increase the number of mod-

els analysed and that these will not be limited to systems biology applications since numerous other fields such as engineering and economics also use ODE equations. This software should facilitate the analysis of models both in SBML and ODE formats.

Another interesting topic that should be explored in greater detail is the structural identifiability analysis of large controlled ODE systems. There has been very little published on this topic and our numerical method can assist in reducing the computational demand associated with these analyses.

The final two future applications stretch beyond the scope of structural identifiability. The general consensus is that the analysis of nonlinear models is difficult. Furthermore, identifiability, observability and controllability are all structural properties of a model that describe the relationships between the state, input, and output variables [21]. Since controllability is the dual problem of observability, a possible application of our method is analysing the controllability of a model. This was alluded to by Stigter *et. al.* in their 2018 conference paper [21] in which they illustrate how to evaluate the local controllability of an ODE model. Similar to the efficiency of the structural identifiability method observed in this thesis, this extension naturally lends itself to the controllability analysis of large ODE models.

Finally, the numerically calculated sensitivity matrix, denoted as $\boldsymbol{S}$, can play a role in practical identifiability analyses [3, 22]. The practical identifiability problem is based on noisy measurements of the dynamics of a measured output over a range of discrete time points. The set of these measured outputs can be denoted by $\boldsymbol{z}$ and defined as [23]:

$$\boldsymbol{z}(t_j) = \boldsymbol{y}(t_j, \boldsymbol{\theta}) + e(t_j), \tag{6.2}$$

where $t_j = 0, \ldots, t_N$ and the error term, $e(t_j)$, is assumed to belong to some distribution. In short, the confidence region of the individual unknown parameters in $\boldsymbol{\theta}$ can be approximated, *a priori*, from the Fisher information matrix [23]. The Fisher information matrix can easily be calculated using the sensitivity matrix as [24, 25]:

$$\boldsymbol{F} = \boldsymbol{S}^T \boldsymbol{S}. \tag{6.3}$$

This matrix is often used in experimental design applications, with the aim of reducing the confidence intervals of the respective parameters. Examples of factors that can be optimised are: 1) the external control inputs, $\boldsymbol{u}(t)$, 2) the experiment's duration, t_N, and 3) sampling times, t_i. The message here is that our method can easily be used in further experimental design applications [26]. Having already calculated a sensitivity matrix, we can easily design other features of an experiment.

REFERENCES

[1] C. Stadtländer, *Systems biology: mathematical modeling and model analysis,* Journal of Biological Dynamics **12**, 11 (2018).

[2] A. F. Villaverde, N. Tsiantis, and J. R. Banga, *Full observability and estimation of unknown inputs, states and parameters of nonlinear biological models,* Journal of The Royal Society Interface **16**, 20190043 (2019).

[3] J. D. Stigter and J. Molenaar, *A fast algorithm to assess local structural identifiability,* Automatica **58**, 118 (2015).

[4] D. L. I. Janzén, L. Bergenholm, M. Jirstrand, J. Parkinson, J. Yates, N. D. Evans, and M. J. Chappell, *Parameter identifiability of fundamental pharmacodynamic models,* Frontiers in Physiology **7**, 590 (2016).

[5] J. Z. Chee, Y. Shiong Chiew, C. P. Tan, and G. Arunachalam, *Identifiability of patient effort respiratory mechanics model,* in *2018 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)* (2018) pp. 48–53.

[6] J. Price, B. Hofmann, V. Silva, M. Nordblad, J. Woodley, and J. Huusom, *Mechanistic modeling of biodiesel production using a liquid lipase formulation,* Biotechnol. Prog. **30**, 1277 (2014).

[7] H. Yu, H. Yue, and P. Halling, *Comprehensive experimental design for chemical engineering processes: A two-layer iterative design approach,* Chemical engineering science **189**, 135 (2018).

[8] M. Anguelova, J. Karlsson, and M. Jirstrand, *Minimal output sets for identifiability,* Mathematical Biosciences **239**, 139 (2012).

[9] M. P. Saccomani, S. Audoly, and L. D'Angió, *Parameter identifiability of nonlinear systems: the role of initial conditions,* Automatica **39**, 619 (2003).

[10] A. F. Villaverde, A. Barreiro, and A. Papachristodoulou, *Structural identifiability of dynamic systems biology models,* PLOS Computational Biology **20**, 1 (2016).

[11] M. J. Chappell and R. N. Gunn, *A procedure for generating locally identifiable reparameterisations of unidentifiable non-linear systems by the similarity transformation approach,* Mathematical Biosciences **148**, 21 (1998).

[12] N. D. Evans and M. J. Chappell, *Extensions to a procedure for generating locally identifiable reparameterisations of unidentifiable systems,* Mathematical Biosciences **168**, 137 (2000).

[13] J. Karlsson, M. Anguelova, and M. Jirstrand, *An efficient method for structural identifiability analysis of large dynamic systems,* IFAC Proceedings Volumes **45**, 941 (2012), 16th IFAC Symposium on System Identification.

[14] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmuller, and J. Timmer, *Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood,* Bioinformatics **25**, 1923–1929 (2009).

[15] J. D. Stigter, D. Joubert, and J. Molenaar, *Observability of complex systems: Finding the gap,* Scientific Reports **7**, 1 (2017).

[16] P. Cumsille, M. Godoy, Z. P. Gerdtzen, and C. Conca, *Parameter estimation and mathematical modeling for the quantitative description of therapy failure due to drug resistance in gastrointestinal stromal tumor metastasis to the liver,* PLOS ONE **14**, 1 (2019).

[17] N. Meshkat, M. Eisenberg, and J. J. DiStefano, *An algorithm for finding globally identifiable parameter combinations of nonlinear ode models using Gröbner Bases,* Mathematical Biosciences **222**, 61 (2009).

[18] G. Bellu, M. P. Saccomani, S. Audoly, and L. D'Angió, *Daisy: A new software tool to test global identifiability of biological and physiological systems,* Computer Methods and Programs in Biomedicine **81**, 52 (2007).

[19] T. Ligon, F. Fröhlich, O. Chis, J. Banga, E. Balsa-Canto, and J. Hasenauer, *Genssi 2.0: multi-experiment structural identifiability analysis of sbml models,* Bioinformatics **34**, 1421–1423 (2018).

[20] N. Meshkat, C. Anderson, and J. J. DiStefano, *Finding identifiable parameter combinations in nonlinear ode models and the rational reparameterization of their input–output equations,* Mathematical Biosciences **233**, 19 (2011).

[21] J. Stigter, L. van Willigenburg, and J. Molenaar, *An efficient method to assess local controllability and observability for non-linear systems,* IFAC-PapersOnLine **51**, 535 (2018), 9th Vienna International Conference on Mathematical Modelling.

[22] R. Brun, M. Kühni, H. Siegrist, W. Gujer, and P. Reichert, *Practical identifiability of asm2d parameters—systematic selection and tuning of parameter subsets,* Water Research **36**, 4113 (2002).

[23] M. Saccomani and K. Thomaseth, *The union between structural and practical identifiability makes strength in reducing oncological model complexity: A case study,* Complexity **2380650** (2018), 10.1155/2018/2380650.

[24] A. Cintrón-Arias, H. Banks, A. Capaldi, and L. A.L., *A sensitivity matrix based methodology for inverse problem formulation,* Journal of Inverse and Ill-posed Problems **17**, 545 (2009).

[25] A. Sinkoe and J. Hahn, *Optimal experimental design for parameter estimation of an il-6 signaling model,* Processes **5** (2017), 10.3390/pr5030049.

[26] N. Krausch, T. Barz, A. Sawatzki, M. Gruber, S. Kamel, P. Neubauer, and M. N. Cruz Bournazou, *Monte carlo simulations for the analysis of non-linear parameter confidence intervals in optimal experimental design,* Frontiers in Bioengineering and Biotechnology **7**, 122 (2019).

ACRONYM LIST

ODE	Ordinary differential equation
SI	Structural identifiability
LTI	Linear time invariant
ORC	Observability rank condition
s.l.i	Structurally locally identifiable
s.g.i	Structurally globally identifiable
SVD	Singular value decomposition
MOS	Minimal output sets
SBML	Systems Biology Markup Language
R.O.T	Rule of thumb

SUMMARY

A fundamental principle of systems biology is that it will always require new technologies to solve challenging biological questions. This precept will continue to drive the development of novel analytical tools, and so the virtuous cycle of biological progress can only exists when experts from different disciplines including biology, chemistry, computer science, engineering, mathematics, physics, and medicine collaborate. General opinion is however that one of the challenges facing the systems biology community is the lag in the development of such technologies.

The research conducted in this thesis aimed to provide systems biologists with a tool that can be used to: 1) assess the structural identifiability of their models, 2) obtain a clear strategy on how to address structural unidentifiability, and 3) preliminary design experiments.

The hybrid property of the algorithm presented here, which combines both numerical and symbolic identifiability calculations, allowed for a significant reduction in the computational demand often associated with identifiability calculations. This enabled us to analyse large ODE models for which certain analyses would be computationally intractable.

In **Chapter 2**, I introduced an iterative identifiability algorithm that could determine minimal sets of outputs that need to be measured to ensure a model's local structural identifiability. I proposed that in the future this algorithm could be used in the preliminary design of experiments and that this would give scientists valuable insight into exactly which sensors they needed to measure to ensure that the unknown parameters of a model could in principle be estimated. I also illustrated how one could potentially reduce the computational demand of the algorithm by randomly omitting states/sensors from the measured output each time it is repeated.

The novelty of the work presented in this chapter is that one could determine the sets of important sensors, even for large systems biology models within a matter of minutes.

In **Chapter 3**, I addressed the topic of structural unidentifiability. I presented a method that could provide theoretical suggestions for the reparameterisation of structurally unidentifiable models. The novelty of this work is that the algorithm allowed for unknown initial conditions to be parameterised and accordingly, repameterisations requiring the transformations of states associated with unidentifiable initial conditions could easily be obtained. In addition, the computational efficiency of the method allowed for the reparameterisation of large ODE models.

In **Chapter 4**, I investigated potential ways in which numerical accuracy could be improved. I preformed a sensitivity analysis of our sensitivity based method to determine which factors where the most influential. I started off by mentioning the pitfalls of the algorithm: 1) accuracy issues due to scaling, 2) problems integrating stiff systems, and 3) the potential effect of number of sensors measured. I concluded that parameter and initial values and the composition of the measured output vector had the most significant impact on numerical results. To this end, I showed that numerical accuracy could be improved by vertically concatenating sensitivity matrices.

In **Chapter 5**, I revisited the topic of structural unidentifiability, this time investigating the role of initial conditions in a model's structural unidentifiability. I showed that the algorithm can be added to a list of 3 software packages capable of detecting problematic initial conditions and that it is capable of detecting these values even for large ODE models. This novel since previous publications on this topic only included small 2 state models as examples.

The method has potential roles in both addressing structural unidentifiability and preliminary experimental design, since the efficiency of the method allows for the detection of problematic initial conditions using an exhaustive search. I showed that problematic initial conditions can be zero or nonzero values or combinations of specific sets of initial and parameters values, and that a model may have multiple sets of these problematic values. A final point raised was that models in steady state might be structurally unidentifiable.

In the General Discussion in **Chapter 6**, I discussed the results obtained in the different chapters and highlighted the important concepts that emerged from each of the chapters. I continued to discuss future work that needs to be done in the context of providing easy-to-use tools to the systems biology community. The main features being the development of an online tool that could be used to analyse models. Ultimately, newly developed technologies need to be advertised and to this end a lot of work remains to be done.

About the Author

Dominique Joubert was born on 12 April 1983 in Pretoria, South Africa. In 2006 she obtained her BSc (Honours) degree in Optometry from the University of Johannesburg. After working as an optometrist, she enrolled in a BIng degree in 2008 and graduated as a Chemical Engineer in 2012. Whilst studying engineering, she developed a keen interest in the field of numerical analysis. As a result, she completed her MSc in Applied Mathematics with distinction from the North West University in 2014. Her master thesis was in financial mathematics in which she calculated a numerical solution to a set of partial differential equations that described the price of American put options under stochastic volatility.

In 2015, Dominique obtained a PhD position at the Mathematical and Statistical Methods Group (Biometris) at Wageningen University under the supervision of Jaap Molenaar and Hans Stigter. Her research was on the topic of structural identifiability. During her PhD, Dominique presented her research at various conferences.

Currently, Dominique is employed at the Mathematical and Statistical Methods Group in Wageningen, where she is developing an application which implements the algorithms presented in her thesis. Dominique's ambition is to continue doing research, develop practical tools for scientists and to transfer her passion for mathematics to the next generation.

LIST OF PUBLICATIONS

D. Joubert, J.D. Stigter and J. Molenaar, *Determining minimal output sets that ensure structural identifiability*, PLOS ONE **13**, 11 (2018).

J.D. Stigter, **D. Joubert** and J. Molenaar, *Observability of Complex Systems: Finding the Gap*, Scientific Reports **7** (2017).

D. Joubert, J.D. Stigter and J. Molenaar, *An efficient procedure to reparameterise structurally unidentifiable models (under review).*

D. Joubert, J.D. Stigter and J. Molenaar, *Assessing the role of initial conditions in the local structural identifiability of large nonlinear dynamical models (under review).*

Conference abstracts

D. Joubert, J.D. Stigter and J. Molenaar, *Determining the minimal output sets that ensure the structural identifiability of a model, Lisabon, European Conference on Mathematical and Theoretical Biology, July, 23-27, 2018 (Oral presentation).*

D. Joubert, J.D. Stigter and J. Molenaar, *An efficient procedure to reparameterise large ODE models, Integrative Collaborative modelling in systems Medicine: Benried, INCOME, October, 13-19, 2018 (Poster presentation).*

D. Joubert, J.D. Stigter and J. Molenaar, *Parameter Identifiability: Large nonlinear models, Lunteren, BioSB, April, 4-5, 2017 (Poster presentation).*

ACKNOWLEDGEMENTS

Completing an important journey in one's life is never done in isolation and so, it is with great humility that I wish to thank the following individuals.

I would like to thank my parents for all their love and support. I do think that a parent's unconditional love gives a child an unimaginable advantage in life. Thank you **Pierre** and **Fiona** for giving me beauty, brains and love that can fill oceans. I only hope that I can be there for you as you have been for me.

Then I would like to thank **Patricia**. Tiesie, words cannot describe my gratitude for all your love and kindness. I thank you for letting me be part of your Dutch family.

A special thank you to the person, who unknowingly, has perhaps had the biggest part to play in the direction my life has taken, **Japie Spoelstra**. I truly say unknowingly, because your influence in my life has never been through any communication, but merely though the manifestation of what you believe in. So thank you for inspiring me to step into the wonderful world of mathematics and for your role in what has now resulted in a doctoral degree.

Thank you also to both my supervisors **Jaap Molenaar** and **Hans Stigter**. Thank you for your guidance, kindness, and utmost professionalism. I have learnt so much from you. Surprisingly perhaps, it is the role you played in my realisation that my decisions can determine the trajectory and purpose of another person's life, that I will keep in highest regard.

I also would like to thank my brother, **Pierre**. Thank you for making the journey to be by my side. I hope you find happiness and the purpose that I know we both have a shared sense of. A thanks also to **Rudie**. Rudie, you are sometimes the only person that can understand what I experience. I hope you too find happiness and that all your academic and personal hopes come to fruition.

Then thank you to my Wageningen family, **Ria**, **Wim**, **Imke** and **Annemerel**. I have been blessed to get to know all of you and I will surely miss our dinners. Ria and Wim, hearing about your great adventures, and Imke, witnessing all your attempts to grow anything that can be grown. Strangely enough, this made me feel right at home.

A warm thank you also to my colleagues at **Biometris**. I find it fascinating that no matter where one goes, mathematicians and statisticians are always of the nicest people you can ever wish to meet. Thank you for the all of your kindness.

Finally, this has been a hard journey. During my absence from South Africa I have lost both grandmothers and my two beloved dachshunds, Woody and Fazakas. I also had to leave behind lovely boy Carlos. And so, I also wish to thank my grandmothers **Aggie** and **Hesie** for their love and influence in my life. In closing, my message is that you never know what influence you have on someone else's life, so make sure it is uplifting.

EDUCATION STATEMENT

PE&RC Training and Education Statement

With the training and education activities listed below the PhD candidate has complicated with the requirements set by the C.T. de Wit Graduate School for Production Ecology and Resource Conservation (PE&RC) which comprises of a minimum total of 32 ECTS (= 22 weeks of activities)

Review of literature (4.5 ECTS)
Uncertainty and identifiability analysis in the life sciences and application to flavour prediction modelling

Writing of project proposal (4.5 ECTS)
Uncertainty and identifiability analysis in the life sciences and application to flavour prediction modelling

Postgraduate courses (7.2 ECTS)
- MATLAB; BIOSB (2015)
- Discovering systems biology principles; BIOSB (2015)
- System identification in the life sciences; DISC (2016)
- Basic statistics; PE&RC (2016)
- Spring congress: mathematics for health; 4TU.AMI (2017)
- Statistical uncertainty analysis of dynamic models; PE&RC (2017)
- Multivariate analysis; PE&RC (2017)

Deficiency, refresh, brush-up courses (3 ECTS)
- Machine learning for spatial data; PE&RC (2018)
- Summer school: data science: data analysis and visualization; Utrecht University (2019)

Competence strengthening / skills courses (6.6 ECTS)
- Philosophy and ethics of food science technology; PE&RC (2016)
- Essentials of scientific writing & presenting; WGS (2017)
- Scientific publishing; WGS (2017)
- Career perspectives; PE&RC (2018)
- Start to teach; PE&RC (2019)
- Presenting with impact; PE&RC (2019)

PE&RC Weekend, PERC Day, and other PERC events (1.2 ECTS)
- PE&RC First years weekend (2015)
- PE&RC Day (2018)

Discussion groups / local seminars or scientific meetings (6.4 ECTS)
- Systems biology and ecology colloquium (2016-2019)
- INtegrative COllaborative modeling in systems MEdicine: INCOME - Hackathon (2018)
- Modelling and simulation discussion group (2019)

International symposia, workshops and conferences (5.7 ECTS)
- ECMTB Conference; oral presentation; Lisabon, Potugal (2018)
- INCOME Conference; poster presentation; Bernried, Germany (2018)
- PhD Symposium; oral presentation; Wageningen, the Netherlands (2019)

Lecturing / supervision of practicals / tutorials (3 ECTS)
- Mathematics 2 (2017-2019)
- Mathematics 3 (2018)

Cover design by D. Joubert and N. Joubert
Thesis layout based on latex style dissertation.cls from TU Delft
Printed by Digiforce, the Netherlands

1. Due to its solid numerical approach, the identifiability algorithm analysed in this thesis allows for the analyses of models that could otherwise *never* be examined. (this thesis)

2. When analysing local observability for a given dynamical system, the *only* role for a Lie algebra, imposed by its associated Lie-bracket operator on the given vector fields, is to *verify* numerical results.
(this thesis)

3. Due to sloppy reporting, many model predictions in the literature are irreproducible and therefore useless.

4. Science will never solve all our problems nor will it answer all our questions.

5. The narrative that quantifies human progress in terms of economic growth is destructive.

6. The solution to global warming lies in the commoditization of oxygen.

7. The current relationship between Africa and China can be described as colonisation version 2.0.

Propositions belong to the thesis entitled,
Structural Identifiability of large Systems Biology models

Dominique Joubert
Wageningen, 21 October 2019