JOURNAL OF ALGEBRAIC STATISTICS
Volume 9, No. 1, 2018, p.30-68
https://jalgstat.library.iit.edu/
ISSN 1309-3452

The geometry of sloppiness

Emilie Dufresne [1,*], Heather A. Harrington[2], Dhruva V. Raman[3]

[1] *School of Mathematical Sciences, University of Nottingham, Nottingham, United Kingdom*
[2] *Mathematical Institute, University of Oxford, Oxford, United Kingdom*
[3] *Engineering Department, University of Cambridge, Cambridge, United Kingdom*

Abstract. The use of mathematical models in the sciences often involves the estimation of unknown parameter values from data. Sloppiness provides information about the uncertainty of this task. In this paper, we develop a precise mathematical foundation for sloppiness as initially introduced and define rigorously key concepts, such as 'model manifold', in relation to concepts of structural identifiability. We redefine sloppiness conceptually as a comparison between the premetric on parameter space induced by measurement noise and a reference metric. This opens up the possibility of alternative quantification of sloppiness, beyond the standard use of the Fisher Information Matrix, which assumes that parameter space is equipped with the usual Euclidean metric and the measurement error is infinitesimal. Applications include parametric statistical models, explicit time dependent models, and ordinary differential equation models.

2000 Mathematics Subject Classifications: 93B30, 62B10, 62F25, 26B10, 08A99, 26E05

Key Words and Phrases: sloppiness, structural identifiability, inference, metric geometry

*Corresponding author.

Email addresses: emilie.dufresne@nottingham.ac.uk (E. Dufresne), harrington@maths.ox.ac.uk (H.A. Harrington), dhruva.raman@eng.cam.ac.uk (D.V. Raman)

1. Introduction

Mathematical models describing physical, biological, and other real-life phenomena contain parameters whose values must be estimated from data. Over the past decade, a powerful framework called "sloppiness" has been developed that relies on Information Geometry [1] to study the uncertainty in this procedure [10, 17, 56, 57, 58, 55]. Although the idea of using the Fisher Information to quantify uncertainty is not new (see for example [20, 45]), the study of sloppiness gives rise to a particular observation about the uncertainty of the procedure and has potential implications beyond parameter estimation. Specifically, sloppiness has enabled advances in the field of systems biology, drawing connections to sensitivity [25, 19, 24], experimental design [4, 37, 25], identifiability [47, 55, 13], robustness [17], and reverse engineering [19, 14]. Sethna, Transtrum and co-authors identified sloppiness as a universal property of highly parameterized mathematical models [61, 56, 54, 25]. More recently a non-local version of sloppiness has emerged, called predictive sloppiness [33]. However, the precise interpretation of sloppiness remains a matter of active discussion in the literature [4, 26, 29].

This paper's main contribution is to serve as a first step towards a unified mathematical framework for sloppiness rooted in algebra and geometry. While our work does not synthesize the entirety of the field, we provide some of the mathematical elements needed to formalize sloppiness as it was initially introduced. We extend the concept beyond time dependent models, in particular, to statistical models. We rigorously define the concepts and building blocks for the theory of sloppiness. Our approach requires techniques from many fields including algebra, geometry, and statistics. We illustrate each new concept with a simple concrete example. The new mathematical foundation we provide for sloppiness is not limited by current computational tools and opens up the way to further work.

Our general setup is a mathematical model M that describes the behavior of a variable $x \in \mathbb{R}^m$ depending on a parameter $p \in P \subseteq \mathbb{R}^r$. Our first step is to explain how each precise choice of perfect data z induces an equivalence relation $\sim_{M,z}$ on the parameter space: two parameters are equivalent if they produce the same perfect data. We then characterize the various concepts of structural identifiability in terms of the equivalence relation $\sim_{M,z}$. Roughly speaking, structural identifiability asks to what extent perfect data determines the value of the parameters. See section 2.

Assume that the perfect data z is a point of $\mathbb{R}^N$ for some N. The second crucial step needed in order to define sloppiness is a map ϕ from parameter space P to data space $\mathbb{R}^N$ giving the perfect data as a function $\phi(p)$ of the parameters known as a "model manifold" in the literature [56, 57, 58, 55], which we rename as a model prediction map. A model prediction map thus induces an injective function on the set of equivalence classes (the set-theoretic quotient $P/\sim_{M,z}$), that is, the equivalence classes can be separated by N functions $P \to \mathbb{R}$. See Section 3.

The next step is to assume that the mathematical model describes the phenomenon we are studying perfectly, but that the "real data" is corrupted by measurement error and the use of finite sample size. That is, we assume that noisy data arises from a random

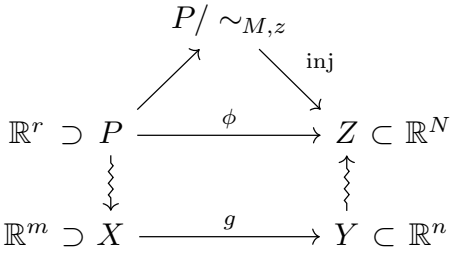
process whose probability distribution then induces a premetric d on the parameter space, via the Kullback-Leibler divergence (see start of Section 4). This premetric d quantifies the proximity between the two parameters in parameter space via the discrepancy between the probability distributions of the noisy data associated to the two parameters.

The aforementioned premetric d has a tractable approximation in the limit of decreasing measurement noise using the Fisher Information Matrix (FIM). In the standard definition, a model is "sloppy" when the condition number of the FIM is large, that is, there are several orders of magnitude between its largest and smallest eigenvalues. Multiscale sloppiness (see [44]) extends this concept to regimes of non-infinitesimal noise.

We conceptually extend the notion of sloppiness to a comparison between the premetric d and a reference metric on parameter space. We demonstrate that using the condition number of the FIM to measure sloppiness at a parameter p_0, as is done in most of the sloppiness literature [10, 17, 56, 57, 58, 55], corresponds to comparing an approximation of d in an infinitesimal neighborhood of p_0 to the standard Euclidean metric on $\mathbb{R}^r \supset P$. Note that considering the entire spectrum of the FIM, as is done newer work in the sloppiness literature (eg, [61]) corresponds to performing a more refined comparison between an approximation of d in an infinitesimal neighborhood of p_0 to the standard Euclidean metric on $\mathbb{R}^r \supset P$. Multiscale sloppiness, which we extend here beyond its original definition [44] for Euclidean parameter space and Gaussian measurement noise, avoids approximating d, and so better reflects the sloppiness of models beyond the infinitesimal scale. Finally, we describe the intimate relationship between sloppiness and practical identifiability, that is, whether generic noisy data allows for bounded confidence regions when performing maximum likelihood estimation. See Section 4.

The following diagram illustrates the main objects discussed in this paper:

M	a mathematical model
x	variable, belongs to $X \subseteq \mathbb{R}^m$
y	observable, belongs to $Y \subseteq \mathbb{R}^n$
ϕ	model prediction map
p	parameter, belongs to $P \subseteq \mathbb{R}^r$
z	data, belongs to $Z \subseteq \mathbb{R}^N$
g	function giving y in terms of x



2. An equivalence relation on parameter space

A mathematical model M describes the behavior of a variable $x \in X \subseteq \mathbb{R}^m$ depending on a parameter $p \in P \subset \mathbb{R}^r$, with measurable output $y = g(x) \in Y \in \mathbb{R}^n$. We further specify a choice of perfect data z produced for the parameter value p. The nature of perfect data will be made clear in the examples discussed throughout the section. We think of the perfect data z as belonging to the wider *data space* Z that encompasses all possible

"real" data. Data space will be defined rigorously in Section 4 when measurement noise comes into play.

An example where the measurable output y differs from $x = (x_1, \ldots, x_n)$ is when only some of the x_i's can be measured (e.g., due to cost or inaccessibility of certain variables). The perfect data is extracted from the measurable output, as illustrated by examples 2.1, 2.5, and 2.7. The behavior of the variable x may also vary in time (and position in space, although this will not be addressed here). In the time dependent case, the perfect data often consists of values of the measurable output y at finitely many timepoints, that is, a *time series*. An alternative choice of perfect data would be the set of all stable steady states. We are also interested in what we will call the *continuous data*, that is, the value of y at all possible timepoints or, equivalently, the function $t \mapsto y(t)$ for t belonging to the full time interval. For a statistical model, the measurable output is the outcome from one instance of a statistical experiment, while a natural choice for perfect data is a probability distribution belonging to the model, or any function or set of functions characterizing this probability distribution.

Given a model M, a choice of perfect data z induces a *model-data equivalence relation* $\sim_{M,z}$ on the parameter space P as follows: two parameters p and p' are equivalent ($p \sim_{M,z} p'$) if and only if fixing the parameter value to p or p' produces the same perfect data. We now provide a more concrete description for a selection of types of mathematical models.

2.1. Finite discrete statistical models

The most straightforward case is when the perfect data is described explicitly as a function of the parameter p. Finite discrete statistical models fall within this group, with the perfect data z being the probability distribution of the possible outcomes depending on the choice of parameter. Such a model is described by a map

$$\begin{aligned} \rho\colon P &\to [0,1]^n \\ p &\mapsto (\rho_1(p), \ldots, \rho_n(p)). \end{aligned}$$

The model-data equivalence relation then coincides with the equivalence relation $\sim_\rho$ induced on P by the map ρ, that is, $p \sim_{M,z} p'$ if and only if $\rho(p) = \rho(p')$.

Example 2.1 (Two biased coins [27]). A person with two biased coins, picks one at random, tosses it and records the result. The person then repeats this three additional times, for a total of four coin tosses. The parameter is $(p_1, p_2, p_3) \in [0,1]^3$, where p_1 is the probability of picking the first coin, p_2 is the probability of obtaining heads when tossing the first coin (that is, the bias of the first coin), and p_3 is the probability of obtaining heads when tossing the second coin. Here, the measurable output is the record of a single instance of the statistical experiment described and perfect data is the probability distribution of the possible outcomes (there are five possibilities). The map giving the model is then

$$\begin{aligned} \rho\colon \quad [0,1]^3 &\to \mathbb{R}^5 \\ (p_1, p_2, p_3) &\mapsto (\rho_0, \rho_1, \rho_2, \rho_3, \rho_4), \end{aligned}$$

where ρ_i is the probability of obtaining heads i times. Explicitly we have

$$\begin{aligned}
\rho_0 &= p_1(1-p_2)^4 + (1-p_1)(1-p_3)^4,\\
\rho_1 &= 4p_1p_2(1-p_2)^3 + 4(1-p_1)p_3(1-p_3)^3,\\
\rho_2 &= 6p_1p_2^2(1-p_2)^2 + 6(1-p_1)p_3^2(1-p_3)^2,\\
\rho_3 &= 4p_1p_2^3(1-p_2) + 4(1-p_1)p_3^3(1-p_3),\\
\rho_4 &= p_1p_2^4 + (1-p_1)p_3^4.
\end{aligned}$$

Two parameters (p_1,p_2,p_3) and (p_1',p_2',p_3') are then equivalent if $\rho(p_1,p_2,p_3) = \rho(p_1',p_2',p_3')$, or equivalently, if $\rho_i(p_1,p_2,p_3) = \rho_i(p_1',p_2',p_3')$ for each i.

We next study the equivalence classes. As we cannot distinguish between the two coins, we will always have $(p_1,p_2,p_3) \sim_{M,z} (1-p_1,p_3,p_2)$, and so the equivalence class of (p_1,p_2,p_3) contains the set $\{(p_1,p_2,p_3),(1-p_1,p_3,p_2)\}$. Furthermore, the equivalence class of (p_1,p_2,p_2) will contain $\{(q_1,p_2,p_2) \mid q_1 \in [0,1]\}$. The equivalence class of $(0,p_2,p_3)$ will contain $\{(0,q_1,p_3) \mid q_1 \in [0,1]\}$ and $\{(1,p_2,q_2) \mid q_2 \in [0,1]\}$.

The ideal $(\rho_i \otimes 1 - 1 \otimes \rho_i \mid i = 0,\ldots,4)$ in $\mathbb{C}[p_1,p_2,p_3] \otimes \mathbb{C}[p_1,p_2,p_3]$ is the ideal cutting out the set-theoretic equivalence relation $\sim_\rho$ on $\mathbb{C}^3$ induced by extending the function ρ to $\mathbb{C}^3$. Indeed, the zero set of this ideal is the set of pairs $((p_1,p_2,p_3),(p_1',p_2',p_3')) \in \mathbb{C}^3 \times \mathbb{C}^3$ such that $(p_1,p_2,p_3) \sim_\rho (p_1',p_2',p_3')$. Using a symbolic computation software, we compute the prime decomposition of its radical and conclude that the equivalence class of $(p_1,p_2,p_3) \in \mathbb{C}^3$ is

$$\begin{cases}
\{(p_1,p_2,p_3),(1-p_1,p_3,p_2)\} & \text{if } p_1 \neq 0,1,1/2\ p_2 \neq p_3,\\
\{(q,p_2,p_2) \mid q \in \mathbb{C}\} & \text{if } p_1 \neq 0,1,1/2\ p_2 = p_3,\\
\{(0,q_1,p_3) \mid q_1 \in \mathbb{C}\} \cup \{(1,p_2,q_2) \mid q_2 \in \mathbb{C}\} & \text{if } p_1 = 0,1,\\
\{(1/2,p_2,p_3)\} & \text{if } p_1 = 1/2.
\end{cases}$$

Therefore, the equivalence classes in $[0,1]^3$ must be contained in the intersections of the above sets with $[0,1]^3$. Thus the equivalence class of $(p_1,p_2,p_3) \in [0,1]^3$ is

$$\begin{cases}
\{(p_1,p_2,p_3),(1-p_1,p_3,p_2)\} & \text{if } p_1 \neq 0,1,1/2\ p_2 \neq p_3,\\
\{(q,p_2,p_2) \mid q \in [0,1]\} & \text{if } p_1 \neq 0,1,1/2\ p_2 = p_3,\\
\{(0,q,p_3) \mid q_1 \in [0,1]\} \cup \{(1,p_2,q) \mid q_2 \in [0,1]\} & \text{if } p_1 = 0,1,\\
\{(1/2,p_2,p_3)\} & \text{if } p_1 = 1/2.
\end{cases}$$

In particular, we obtain a stratification of parameter space as shown in Fig. 1.

We remark that almost all equivalence classes have dimension zero, although some equivalence classes have dimension one. As the points with zero-dimensional equivalence classes form a dense open subset of parameter space, we say that the dimension of an equivalence class is generically zero. Note that since all these zero-dimensional equivalence classes have size two, we say that the equivalence classes are generically of size two. ◁

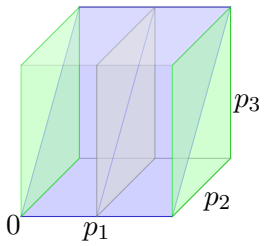


Figure 1: Stratification of parameter space for the two biased coins example. Blue: $\{(p_1, p_2, p_3) \mid p_1 \neq 0, 1, 1/2\ p_2 = p_3\}$ Green: $\{(p_1, p_2, p_3) \mid p_1 = 0, 1, 1/2\}$, Grey: $\{(p_1, p_2, p_3) \mid p_1 = 1/2\}$ the rest of the cube (interior and faces) is the generic part $\{(p_1, p_2, p_3) \mid p_1 \neq 0, 1,\ p_2 \neq p_3\}$.

2.2. time dependent models and the $2r + 1$ result

Let M be an explicit time dependent model with measurable output x. That is, the behavior of the variable x is given by the map

$$\begin{aligned} \rho\colon P \times \mathbb{R}_{\geq 0} &\to \mathbb{R}^m \\ (p, t) &\mapsto x(p, t), \end{aligned}$$

and x can be measured at any time t. Perfect time series data produced by the parameter p will be $(x(p, t_1), \ldots, x(p, t_N))$, where $0 \leq t_1 < \cdots < t_N \in \mathbb{R}_{\geq 0}$ are timepoints. We denote the corresponding model-data equivalence relation on P by $\sim_{M,t_1,\ldots,t_N}$. The continuous data is the map $\mathbb{R}_{\geq 0} \to \mathbb{R}^m$ given by $t \mapsto x(p, t)$. We denote the equivalence relation induced by the continuous data on P by $\sim_{M,\infty}$.

We particularly consider ODE systems with time series data. For such a model M, the behavior of the variable x is described by a system of ordinary differential equations depending on the parameter $p \in P$ with some initial conditions:

$$\begin{aligned} \dot{x} \quad &= f(p, x) \\ x(0) &= x_0. \end{aligned} \tag{1}$$

When initial conditions are known, or we do not wish to estimate them, they are not considered as components of the parameter. The measurable output is $y = g(x)$, and perfect data is then $(y(t_1), \ldots, y(t_N)) \in \mathbb{R}^{Nn}$ for $0 \leq t_1 < \cdots < t_N \in \mathbb{R}_{\geq 0}$. The continuous data is given by the function $\mathbb{R}_{\geq 0} \to \mathbb{R}^n,\ t \mapsto y(t)$, which supposes that a solution to the given ODE system exists, a valid assumption in the real-analytic case.

The key result when working with time dependent models with time series data is the $2r + 1$ result of Sontag [52, Theorem 1], which implies that there is a single "global" model-data equivalence relation: the equivalence relation $\sim_{M,\infty}$ induced by the continuous data. Precisely, we suppose that the model M is real-analytic, that is, either an explicit time-dependent model given by a real-analytic map or an ODE system as in (1) with f a real-analytic function. We additionally assume that the variable x, the parameter p, and the time variable t belong to real-analytic manifolds. If we suppose that P is a real-analytic manifold of dimension r, then for $N \geq 2r + 1$ and a generic choice of timepoints $t_1, \ldots, t_N$ the equivalence relation $\sim_{M,t_1,\ldots,t_N}$ coincides with the equivalence relation $\sim_{M,\infty}$.

An important consequence of the $2r+1$ result [52] is that for real-analytic time-dependent models with time series data, the model equivalence relation is a *global* structural property of the model, and one need not specify which exact timepoints are used.

Remark 2.2. Note that in many applications the variable x belongs to the real positive orthant, which is indeed a real-analytic manifold. The condition on the time variable can be relaxed to include closed and partially closed time intervals.

Remark 2.3. A choice of N timepoints corresponds to a choice of a point in the real analytic manifold $T := \{(t_1, \ldots, t_N) \in \mathbb{R}_{\geq 0} \mid t_i < t_{i+1}\}$. The use of the word "generic" in the statement means that there can be choices of N timepoints that will not induce the equivalence relation $\sim_{M,\infty}$, but that these choices of timepoints will belong to a small subset of T, so small that its complement contains an open dense subset of T.

In cases where no results like the $2r+1$ result [52] hold, there is no "global" equivalence relation. Therefore, a finite number of measurements will never induce the same equivalence relation on parameter space as the continuous data. In other words, by taking more and more measurements we could obtain an increasingly fine equivalence relation without ever converging to $\sim_{M,\infty}$.

Example 2.4 (A model for which the $2r+1$ result does not hold, cf [52, Section 2.3]). The model, while artificial, is an explicit time dependent model given by the map:

$$\begin{aligned} \rho\colon \mathbb{R}_{>0} \times \mathbb{R}_{\geq 0} &\to \mathbb{R} \\ (p, t) &\mapsto \gamma(p - t), \end{aligned}$$

where $\gamma\colon \mathbb{R} \to \mathbb{R}$ is a C^∞ map that is $e^{1/s}$ for $s < 0$ and zero for $s \geq 0$. Suppose for a contradiction that evaluating at timepoints $t_1, \ldots, t_N$ induces the same equivalence relation on $\mathbb{R}_{>0}$ as taking the perfect data to be the maps $t \mapsto \rho(p, t)$. Take $p_1 > p_2 \geq t_N$, it follows that $\rho(p_1, t_i) = 0 = \rho(p_2, t_i)$ for each $i = 1, \ldots, N$. On the other hand, we will have $\rho(p_1, {}^{p_1+p_2}/_2) = 0 \neq \rho(p_2, {}^{p_1+p_2}/_2)$, and so we have a contradiction. ◁

Example 2.5 (Fitting points to a line). This example is motivated by one of the examples found on the webpage of Sethna dedicated to sloppiness [50]. We consider an explicit time dependent model where the variable x changes linearly in time:

$$x(t) = a_0 + a_1 t,$$

that is, x is given as a polynomial function in t depending on the parameter $(a_0, a_1) \in \mathbb{R}^2$. Hence, by the $2r+1$ result [52], taking the perfect data to be the measurement at $2 \cdot 2 + 1 = 5$ sufficiently general time points induces the same equivalence relation as taking the perfect data as the continuous function $t \mapsto a_0 + a_1 t$. In fact, taking measurements at two timepoints will suffice, since there is exactly one line going through any two given points.

We have that $(a_0, a_1) \sim_{M,\infty} (b_0, b_1)$ if and only if

$$a_0 + a_1 t = b_0 + b_1 t, \text{ for all } t \in \mathbb{R}_{\geq 0}.$$

It follows that $a_0 = b_0$ (taking $t = 0$), and then $a_1 = b_1$ (taking $t = 1$), thus $[(a_o, a_1)]_{M,\infty} = \{(a_0, a_1)\}$. Naturally, this coincides with the equivalence classes obtained

with taking the perfect data to be noiseless measurements at $t = 0$ and $t = 1$, that is, $(x(0), x(1)) = (a_0, a_0 + a_1)$. ◁

Example 2.6 (Sum of exponentials). The sum of exponentials model for exponential decay, widely studied in the sloppiness literature [56, 57, 58], is an explicit time dependent model given by the function

$$\begin{aligned}\rho\colon \mathbb{R}^2_{\geq 0} \times \mathbb{R}_{\geq 0} &\to \mathbb{R} \\ (a, b, t) &\mapsto e^{-at} + e^{-bt}.\end{aligned}$$

By the $2r+1$ result [52], the time series $(e^{-at_1}+e^{-bt_1}, \ldots, e^{-at_5}+e^{-bt_5})$ with $(t_1, \ldots, t_5)$ generic induces the same equivalence relation on the parameter space $\mathbb{R}^2_{\geq 0}$ as the continuous data. This model is clearly non-identifiable. Indeed, for any $a, b \in \mathbb{R}_{\geq 0}$, the parameters (a, b) and (b, a) yield the same continuous data since $e^{-at} + e^{-bt} = e^{-bt} + e^{-at}$ for all t. It follows that the equivalence class of a parameter (a, b) will contain the set $\{(a, b), (b, a)\}$.

Suppose $(a, b) \sim_{M,t_1,t_2} (a', b')$ where $t_1 \neq t_2$ are positive real numbers, thus

$$\begin{aligned}e^{-at_1} + e^{-bt_1} &= e^{-a't_1} + e^{-b't_1}, \\ e^{-at_2} + e^{-bt_2} &= e^{-a't_2} + e^{-b't_2}.\end{aligned}$$

We can reduce it to the case $t_1 = 1$, $t_2 = 2$ by rescaling the time variable via the substitution $t \mapsto {}^{(t+t_2-2t_1)}/_{(t_2-t_1)}$ in ρ. Simplifying further with the substitution $x = e^{-a}, y = e^{-b}, u = e^{-a'}, v = e^{-b'}$, the equation becomes:

$$\begin{aligned}x + y &= u + v \\ x^2 + y^2 &= u^2 + v^2.\end{aligned}$$

It is then easy to see that the only solutions (u, v) to this system are $(u, v) = (x, y)$ or $(u, v) = (y, x)$. As the exponential function is injective it follows that $(a', b') = (a, b)$ or $(a', b') = (b, a)$.

Therefore, the equivalence class of a parameter (a, b) is

$$\begin{cases}\{(a, b), (b, a)\}, & \text{if } a \neq b, \\ \{(a, a)\}, & \text{if } a = b.\end{cases}$$

◁

Example 2.7 (An ODE system with a solution). We consider the ODE system with variable $(x_1, x_2) \in \mathbb{R}^2_{\geq 0}$ and parameter $(p_1, p_2) \in \mathbb{R}^2_{>0}$ given by

$$\begin{aligned}\dot{x}_1 &= -p_1 x_1 \\ \dot{x}_2 &= p_1 x_1 - p_2 x_2\end{aligned}$$

with known initial conditions $x_1(0) = c_1$ and $x_2(0) = 0$, and observable output (x_1, x_2). Set $U := \{(p_1, p_2) \mid p_1 \neq p_2\}$. For $(p_1, p_2) \in U$, a solution to this system is given by

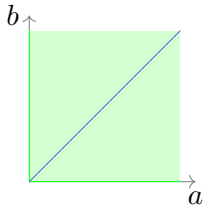


Figure 2: Parameter space of sum of exponential example. Green: $\{(a,b) \in \mathbb{R}^2_{\geq 0} \mid a \neq b\}$, Blue: $\{(a,a) \mid a \in \mathbb{R}_{\geq 0}\}$

$$\begin{aligned} x_1(t) &= c_1 e^{-p_1 t} \\ x_2(t) &= \frac{c_1 p_1}{(p_2 - p_1)} \left(e^{-p_1 t} - e^{-p_2 t}\right). \end{aligned}$$

When $p_1 = p_2$, the ODE system becomes

$$\begin{aligned} \dot{x}_1 &= -p_1 x_1 \\ \dot{x}_2 &= p_1(x_1 - x_2), \end{aligned}$$

and a solution is given by

$$\begin{aligned} x_1(t) &= c_1 e^{-p_1 t} \\ x_2(t) &= c_1 p_1 t e^{-p_1 t}. \end{aligned}$$

The $2r+1$ result [52] implies that, for general $(t_1, \ldots, t_5)$, the time series data $(x_1(t_1), x_2(t_1), \ldots, x_1(t_5), x_2(t_5))$ induces the same equivalence relation on the parameter space $\mathbb{R}^2_{\geq 0}$ as the continuous data. As in the previous example, we will show that this can be achieved by taking a time series with two distinct nonzero time points. We can again reduce to the case $t_1 = 1, t_2 = 2$. Suppose that (p_1, p_2) and (p_1', p_2') are two parameters that produce the same perfect data. The first case we consider is when they both belong to U, then we have

$$\begin{aligned} c_1 e^{-p_1} &= c_1 e^{-p_1'}, \\ c_1 e^{-2p_1} &= c_1 e^{-2p_1'}, \\ \frac{c_1 p_1}{(p_2 - p_1)}(e^{-p_1} - e^{-p_2}) &= \frac{c_1 p_1'}{(p_2' - p_1')}(e^{-p_1'} - e^{-p_2'}), \\ \frac{c_1 p_1}{(p_2 - p_1)}(e^{-2p_1} - e^{-2p_2}) &= \frac{c_1 p_1'}{(p_2' - p_1')}(e^{-2p_1'} - e^{-2p_2'}). \end{aligned}$$

The first equation implies that $p_1 = p_1'$ since $c_1 \neq 0$ and the exponential function is injective. Using the last two equations we find that we have

$$e^{-p_1} + e^{-p_2} = \frac{\frac{c_1 p_1}{(p_2 - p_1)}(e^{-2p_1} - e^{-2p_2})}{\frac{c_1 p_1}{(p_2 - p_1)}(e^{-p_1} - e^{-p_2})} = \frac{\frac{c_1 p_1'}{(p_2' - p_1')}(e^{-2p_1'} - e^{-2p_2'})}{\frac{c_1 p_1'}{(p_2' - p_1')}(e^{-p_1'} - e^{-p_2'})} = e^{-p_1'} + e^{-p_2'},$$

And since $p_1 = p_1'$, it follows that $p_2 = p_2'$. Next, if we suppose that neither belongs to U, that is, (p_1, p_1) and (p_1', p_1') produce the same perfect data, we then have

$$\begin{aligned} c_1 e^{-p_1} &= c_1 e^{-p_1'}, \\ c_1 e^{-2p_1} &= c_1 e^{-2p_1'}, \\ c_1 p_1 e^{-p_1} &= c_1 p_1' e^{-p_1'}, \\ 2c_1 p_1 e^{-2p_1} &= 2c_1 p_1' e^{-2p_1'}. \end{aligned}$$

The first equation already implies that $p_1 = p_1'$. Finally, we suppose that one parameter is in U and the other is not, that is, (p_1, p_2) with $p_1 \neq p_2$ and (p_1', p_1') produce the same perfect data. We then have

$$\begin{aligned} c_1 e^{-p_1} &= c_1 e^{-p_1'} \\ c_1 e^{-p_1^2} &= c_1 e^{-p_1'^2} \\ \frac{c_1 p_1}{(p_2 - p_1)}(e^{-p_1} - e^{-p_2}) &= c_1 p_1' e^{-p_1'} \\ \frac{c_1 p_1}{(p_2 - p_1)}(e^{-2p_1} - e^{-2p_2}) &= 2c_1 p_1' e^{-2p_1'}. \end{aligned}$$

The first two equations imply that $p_1 = p_1'$ and so the last two equations become

$$\begin{aligned} \frac{c_1 p_1}{(p_2 - p_1)}(e^{-p_1} - e^{-p_2}) &= c_1 p_1 e^{-p_1} \\ \frac{c_1 p_1}{(p_2 - p_1)}(e^{-2p_1} - e^{-2p_2}) &= 2c_1 p_1 e^{-2p_1}. \end{aligned}$$

If $p_1 = 0$, then p_2 is not further constrained. If $p_1 \neq 0$, the equations simplify to

$$\begin{aligned} \frac{1}{(p_2 - p_1)}(e^{-p_1} - e^{-p_2}) &= e^{-p_1} \\ \frac{1}{(p_2 - p_1)}(e^{-2p_1} - e^{-2p_2}) &= 2e^{-2p_1}, \end{aligned}$$

and so

$$e^{-p_1} + e^{-p_2} = \frac{\frac{1}{(p_2 - p_1)}(e^{-2p_1} - e^{-2p_2})}{\frac{1}{(p_2 - p_1)}(e^{-p_1} - e^{-p_2})} = \frac{2e^{-2p_1}}{e^{-p_1}} = 2e^{-p_1}.$$

But this implies that $p_1 = p_2$, a contradiction. Hence, the third case was not possible in the first place.

We conclude that the equivalence class of the parameter $(p_1, p_2) \in P$ is

$$\begin{cases} \{(p_1, p_2)\} & \text{if } p_1 \neq 0, \\ \{(0, q) \mid q \in \mathbb{R}_{\geq 0}\} & \text{if } p_1 = 0. \end{cases}$$

◁

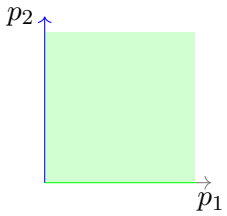


Figure 3: Parameter space of ODE example. Blue: $p_1 = 0$, and Green: $\{(p_1, p_2) \in \mathbb{R}^2_{\geq 0} \mid p_1 \neq 0\}$.

Example 2.7 is an exception. In general one cannot so easily find an exact solution to an ODE system. Nevertheless, describing the equivalence classes can still be possible. Indeed, there are various approaches to building what is called in the literature an exhaustive summary (see for example [41]). An *exhaustive summary* is simply a (not necessarily finite) collection E of functions $P \to \mathbb{R}$ that makes the model-data equivalence relation effective, that is, $p \sim_M p'$ if and only if $f(p) = f(p')$ for all $f \in E$. The differential algebra approach, introduced by Ljung and Glad [36] and Ollivier [42], relies on using exhaustive summaries. For an ODE system with time series data given by rational functions, one derives an input-output equation whose coefficients (once normalized so that the first term is one) provide an exhaustive summary for a dense open subset of parameter space.

Additional details on exhaustive summaries are given by Ollivier [42] and Meshkat et al. [39], and software is available for computing input-output equations [5]. Exhaustive summaries are useful for determining identifiability (subsequently defined) and finding identifiable parameter combinations.

2.3. Structural Identifiability

We formulate a definition of structural identifiability in terms of the model-data equivalence relation defined at the beginning of this section. We base our rigorous understanding of the various flavors of identifiability in Sullivant's in-progress book on Algebraic Statistics [53] and Di Stefano III's book on Systems Biology [28].

Definition 2.8 (Structural Identifiability)**.** Let (M, z) be a mathematical model with a choice of perfect data z inducing an equivalence relation $\sim_{M,z}$ on the parameter space P.

- The pair (M, z) is *globally identifiable* if every equivalence class consists of a single element.
- The pair (M, z) is *generically identifiable* if for almost all $p \in P$, the equivalence class of p consist of a single element.
- The pair (M, z) is *locally identifiable* if for almost all $p \in P$, the equivalence class of p has no accumulation points.
- The pair (M, z) is *non-identifiable* if at least one equivalence class contains more than one element.

- The pair (M, z) is *generically non-identifiable* if for almost all $p \in P$, the equivalence class of p has accumulation points (or are positive dimensional).

Remark 2.9. In the definition above, "almost all" is used to mean that the property holds on a dense open subset of parameter space with respect to the usual Euclidean topology on $\mathbb{R}^r \supset P$. Recall also that $q \in Q \subseteq P \subseteq \mathbb{R}^r$ is an accumulation point of Q if every open neighborhood of p contains infinitely many elements of Q.

Remark 2.10. In the ODE systems literature, where local identifiability is the main concern, "non-identifiable" is often used to mean what we have called "generically non-identifiable".

Example 2.11 (The sum of exponentials). We revisit Example 2.6 where we computed the equivalence class of any parameter $(a, b) \in \mathbb{R}_{\geq 0}$. We found that $[(a, b)] = \{(a, b), (b, a)\}$, where $[(a, b)]$ denotes the set of parameters equivalent to (a, b). It follows that this model is not globally identifiable, and so non-identifiable. This model is locally identifiable since every equivalence class has size at most 2. ◁

Example 2.12 (Two biased coins). We revisit the model considered in Example 2.1. We showed that the equivalence class of a parameter $(p_1, p_2, p_3) \in [0, 1]^3$ is

$$\begin{array}{ll} \{(p_1, p_2, p_3), (1 - p_1, p_3, p_2)\} & \text{if } p_1 \neq 0, 1, 1/2 \; p_2 \neq p_3 \\ \{(q, p_2, p_2) \mid q \in [0, 1]\} & \text{if } p_2 = p_3 \\ \{(0, q, p_3) \mid q \in [0, 1]\} \cup \{(1, p_2, q) \mid q \in [0, 1]\} & \text{if } p_1 = 0, 1, 1/2 \; p_2 \neq p_3, \\ \{(1/2, p_2, p_3)\} & \text{if } p_1 = 1/2. \end{array}$$

This model is not globally identifiable (in fact no equivalence class is a singleton), but it is locally identifiable. Indeed, the equivalence classes have size two for almost all values of the parameter; only the parameters in the 2-dimensional subset $\{(p_1, p_2, p_3) \mid p_1(p_1 - 1)(p_2 - p_3) = 0\}$ have positive dimensional equivalence classes. ◁

Example 2.13 (Fitting points to a line). We revisit the model discussed in Example 2.5. We saw that $[(a_0, a_1)] = \{(a_0, a_1)\}$ for all possible values of the parameter, therefore this model is globally identifiable. ◁

Example 2.14 (An ODE system with an exact solution). For the model studied in Example 2.7, the equivalence class of a parameter $p = (p_1, p_2) \in P = \mathbb{R}^2_{\geq 0}$ is

$$\begin{array}{ll} \{(p_1, p_2)\} & \text{if } p_1 \neq 0, \\ \{(0, q) \mid q \in \mathbb{R}_{\geq 0}\} & \text{if } p_1 = 0. \end{array}$$

As some equivalence classes are infinite, this model is not globally identifiable, but it is generically identifiable. Indeed, the equivalence classes of parameters belonging to the dense open subset $\{(p_1, p_2) \in P \mid p_1 \neq 0\}$ have size 1. ◁

Example 2.15 (A nonlinear ODE model, see [38, Example 6] and [40, Example 5]). We now consider a model given by an ODE system with time series data and describes the

behavior of a variable (x_1, x_2) depending on a 5-dimensional parameter $(p_1, p_2, p_3, p_4, p_5)$ with measurable output $y = x_1$. The ODE system is given by:

$$\begin{aligned}\dot{x}_1 =& p_1 x_1 - p_2 x_1 x_2\\ \dot{x}_2 =& p_3 x_2(1 - p_4 x_2) + p_5 x_1 x_2\end{aligned}$$

The differential algebra method produces an exhaustive summary

$$\phi_1 = \frac{p_3 p_4}{p_2} - 1,\ \phi_2 = \frac{-2p_1 p_3 p_4}{p_2} - p_3,\ \phi_3 = -p_5,\ \phi_4 = \frac{p_1^2 p_3 p_4}{p_2} + p_1 p_3,\ \phi_5 = p_1 p_5.$$

That is, there is a dense open subset $U \subseteq P$ on which the model-data equivalence relation coincides with the equivalence relation given by the map

$$\begin{aligned}\phi\colon \quad & U \to \mathbb{R}^4\\ & (p_1, p_2, p_3, p_4, p_5) \mapsto (\phi_1, \phi_2, \phi_3, \phi_4, \phi_5)\end{aligned}$$

We may take U to be the set of parameters such that all p_2, p_5 and $2p_2 + p_2 p_3 + p_1 p_2 - 4p_1 p_3 p_4$ are nonzero. Then for $(p_1, p_2, p_3, p_4, p_5) \in U$, we have

$$p_1 = -\frac{\phi_5}{2\phi_3}, \qquad p_3 = -2 - \phi_2 - \frac{2\phi_1\phi_5}{\phi_3}, \tag{2}$$

$$\frac{p_4}{p_2} = \frac{\phi_3(1+\phi_1)}{-2\phi_3 - \phi_2\phi_3 - 2\phi_1\phi_5}, \qquad p_5 = \phi_5. \tag{3}$$

Let $\rho\colon U \to \mathbb{R}^4$ be the map given by $(p_1, p_2, p_3, p_4, p_5) \mapsto (p_1, p_3, p_4/p_2, p_5)$. The map ϕ factors through ρ, and the formulas (2),(3) above provide an inverse for the induced function $\phi\colon \rho(U) \to \phi(U)$, and so in particular this function is bijective. It follows that for $(p_1, p_2, p_3, p_4, p_5) \in U$ the function ρ determines the model-data equivalence relation. Therefore, the equivalence class of $(p_1, p_2, p_3, p_4, p_5) \in U$ is

$$\left\{ \left(p_1, q_1, p_3, \frac{p_4}{p_2} \cdot q \right) \;\middle|\; q \in \mathbb{R} \right\}.$$

Hence, all parameters in U have a 1-dimensional equivalence class and we conclude that the model is generically non-identifiable. ◁

The main strategy we employed in the above examples was to construct a map $\phi\colon P \to \mathbb{R}^N$ for some N, such that $p \sim_{M,z} p'$ if and only if $\phi(p) = \phi(p')$, that is, a map making the equivalence relation $p \sim_{M,z} p'$ effective. The model we considered was given in this way, or we evaluated an explicit time dependent model (or a solution to an ODE model) at finitely many timepoints, or else we used an alternative method to obtain an exhaustive summary and thus such a map. When the model-data equivalence relation can be made effective via a differentiable map $f\colon P \to \mathbb{R}^N$, that is, when we can find f such that $\sim_{M,z} = \sim_f$, it is also possible to determine the local identifiability of the model by looking at the Jacobian of f. The model is locally identifiable if and only if the Jacobian

of f has full rank for generic values of p. Indeed, this is an immediate consequence of the Inverse Function Theorem. This method is regularly employed in algebraic statistics when considering specific models (see for example [53, Proposition 15.1.7]).

In the case of ODE systems for which we do not have a solution and are unable to obtain an exhaustive summary, there are computational methods for establishing the (local) identifiability, see e.g. [41], [47] for a survey of the techniques available.

3. Model Predictions

In this section we provide a rigorous definition and a more mathematically correct name for "model manifold", a geometric object that takes center stage in the sloppiness literature [56, 57, 58, 55].

Definition 3.1. Let M be a mathematical model with parameter space P and a choice of perfect data. Suppose that the perfect data produced for each parameter value $p \in P$ is a point of $\mathbb{R}^N$ for some N. A *model prediction map* is a map $\phi\colon P \to \mathbb{R}^N$ that expresses the perfect data produced for the parameter value p as a function $\phi(p)$.

A model prediction map is a geometric realization of the quotient $P/\!\sim_{M,z}$ in the sense that it factors through the set-theoretic quotient $P \to P/\sim_{M,z}$ in such way that the induced map $\overline{\phi}\colon P/\!\sim_{M,z} \to \mathbb{R}^N$ is injective.

A model prediction map is meant to be more than just a map making the model-data equivalence relation effective: we want to use this map to perform parameter estimation by finding the nearest model prediction (in the image of ϕ) to a given noisy data point (in the data space, possibly off the image of ϕ).

Remark 3.2. The sloppiness literature uses the term "model manifold" for the image of a model prediction map [56, 57, 58, 55]. Although in general the image of ϕ is not a manifold as such, using the term manifold has the benefit of bringing into focus the geometric structure of mathematical models.

Remark 3.3. Note that we do not require a model prediction map to satisfy the universal property of a categorical quotient, that is, we do not require that any map that is constant on the equivalence class factors through ϕ.

Each fiber of ϕ is a single equivalence class. As a consequence, when there is a model prediction map, then $\sim_{M,z} = \ \sim_{\phi}$, that is, the model-data equivalence relation coincides with the equivalence relation induced by ϕ. Therefore, identifiability can be characterized in terms of model prediction maps:

Proposition 3.4. Let M be a mathematical model and suppose there is a model prediction map $\phi\colon P \to \mathbb{R}^N$ for some $N > 0$. Then

- The pair (M, ϕ) is *globally identifiable* if ϕ is injective.
- The pair (M, ϕ) is *generically identifiable* if ϕ is generically injective.

- The pair (M, ϕ) is *locally identifiable* if almost all non-empty fibers of ϕ have no accumulation points.
- The pair (M, ϕ) is *non-identifiable* if ϕ is not injective.
- The pair (M, ϕ) is *generically non-identifiable* if almost all non-empty fibers of ϕ have accumulation points.

In some situations, it may be possible to construct a model prediction map only on a dense open subset of parameter space. A subset $E \subseteq P$ is $\sim_{M,z}$*-stable* if $p \in E$ and $p' \sim_{M,z} p$ implies $p' \in E$, that is, E is the union of equivalence classes.

Definition 3.5. A *generic model prediction map* is a model prediction map $\varphi\colon U \to \mathbb{R}^N$ that is defined on a $\sim_{M,z}$-stable dense open subset $U \subseteq P$ of parameter space.

We will use the notation $\varphi\colon P \dashrightarrow \mathbb{R}^N$ borrowed from rational maps in the algebraic category to denote generic model prediction map when the exact domain of definition is unknown or not important. Three of the above notions of identifiability can be rephrased in terms of generic model prediction maps:

Proposition 3.6. Let M be a mathematical model and suppose there is a generic model prediction map $\varphi\colon P \dashrightarrow \mathbb{R}^N$ for some $N > 0$. Then

- The pair (M, φ) is *generically identifiable* if φ is injective on its domain of definition.
- The pair (M, φ) is *locally identifiable* if almost all non-empty fibers of φ have no accumulation points.
- The pair (M, φ) is *generically non-identifiable* if almost all non-empty fibers of φ have accumulation points.

In the algebraic category, we have an additional notion of identifiability:

Definition 3.7 (Rational Identifiability)**.** Let (M, ϕ) (resp. (M, φ)) be a mathematical model with and algebraic model prediction map defined over $\mathbb{R}$ (resp. a generic model prediction map given by a rational map with real coefficients). We say that (M, ϕ) (resp. (M, φ)) is *rationally identifiable* if and only if each parameter p_j can be written as a rational function of the ϕ_i's (resp. the φ_i's), or equivalently if the fields of rational functions are equal: $\mathbb{R}(p_1, \ldots, p_r) = \mathbb{R}(\phi_1, \ldots, \phi_n)$ (resp. $\mathbb{R}(p_1, \ldots, p_r) = \mathbb{R}(\varphi_1, \ldots, \varphi_n)$).

Note that rational identifiability implies generic identifiability. The implication is strict because we are working over a non-algebraically closed field (i.e. $\mathbb{R}$).

Example 3.8 (An example of global identifiability, but not rational identifiability). Consider the model M with model prediction map $\phi : \mathbb{R} \to \mathbb{R}$ defined on the parameter space $\mathbb{R}$ by $p \mapsto p^3 + p$. First, we show that M is globally identifiable. Let a and b be two real numbers such that $a^3 + a = b^3 + b$. We can rewrite $a^3 + a = b^3 + b$ as $(a-b)(a^2+ab+b^2+1) = 0$. The polynomial function $a^2 + ab + b^2 + 1$ has no real zeros, since for any given $b \in \mathbb{R}$, it is a polynomial of degree 2 in a with discriminant $-3b^2 - 4 < 0$. It follows that $a = b$, and so the model is globally identifiable. As x is not a rational function of $x^3 + x$, (M, ϕ) is not rationally identifiable. ◁

The case of finite discrete parametric statistical models is again the simplest case, since the parameterization map is a model prediction map. For the two biased coin model studied in Examples 2.1 and 2.12, the map ϕ is a model prediction map. It is possible to have non-isomorphic sets of model predictions, and also, as in the following example, we may have model prediction maps belonging to different categories (real-analytic vs algebraic).

Example 3.9 (Gaussian Mixtures). We consider the mixture of two 1-dimensional Gaussians, a model that can be used to describe the behavior of one measurement we make on individuals belonging to two populations. The model goes back to Pearson in 1894 who developed the methods of moments while studying crabs in the Bay of Naples. We follow the treatment by Améndola, Faugère and Sturmfels [2]. The parameter is 5-dimensional: $(\lambda, \mu, \sigma, \nu, \tau) \in [0,1] \times \mathbb{R} \times \mathbb{R}_{\geq 0} \times \mathbb{R} \times \mathbb{R}_{\geq 0} =: P$. The mixing parameter λ gives the proportion of the first population, the remaining four coordinate parameters are the means and variances of the two Gaussian distributions: μ, σ and ν, τ. Note that this model is at best locally identifiable. Indeed, since we cannot tell to which population an individual belongs, the parameters $(\lambda, \mu, \sigma, \nu, \tau)$ and $(1-\lambda, \nu, \tau, \mu, \sigma)$ will induce the same probability distribution (that is, the same perfect data) and so we will have $[(\lambda, \mu, \sigma, \nu, \tau)] \supseteq \{(\lambda, \mu, \sigma, \nu, \tau), (1-\lambda, \nu, \tau, \mu, \sigma)\}$, that is, the equivalence class of a parameter includes its orbit under an affine action of the symmetric group on two elements. It follows that generic equivalence classes will have size at least 2. Non-generic special cases will include the case where both populations have the same behavior, that is, $(\mu, \sigma) = (\nu, \tau)$, and the case where only one population is actually present, that is, $\lambda = 0$ or $\lambda = 1$. In these cases the equivalence class of a parameter contains certain subsets as follows:

$$\begin{aligned}
[(\lambda, \mu, \sigma, \mu, \sigma)] &\supseteq \{(q, \mu, \sigma, \mu, \sigma) \mid q \in [0,1]\} && \text{if } (\mu, \sigma) = (\nu, \tau) \\
[(0, \mu, \sigma, \nu, \tau)] &\supseteq \{(0, q_1, q_2, \nu, \tau), (1, \nu, \tau, q_1, q_2) \mid q_1 \in \mathbb{R},\ q_2 \in \mathbb{R}_{\geq 0}\} && \text{if } \lambda = 0 \\
[(1, \mu, \sigma, \nu, \tau)] &\supseteq \{(1, \mu, \sigma, q_1, q_2), (0, q_1, q_2, \mu, \sigma) \mid q_1 \in \mathbb{R},\ q_2 \in \mathbb{R}_{\geq 0}\} && \text{if } \lambda = 1
\end{aligned}$$

In particular, some non-generic equivalence classes will be 1 and 2-dimensional.

As well as a cumulative distribution function $F(x)$, this model has both a probability density function $f(x)$ and a moment generating function $M(t)$; either characterizes the model. The probability density function is the map

$$\begin{aligned}
f \colon \mathbb{R} \times P &\to \mathbb{R} \\
x &\mapsto \lambda \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \right) + (1-\lambda) \left(\frac{1}{\tau\sqrt{2\pi}} e^{-\frac{(x-\nu)^2}{2\tau^2}} \right),
\end{aligned}$$

the cumulative distribution function is the map

$$\begin{aligned}
F \colon \mathbb{R} \times P &\to \mathbb{R} \\
x &\mapsto \lambda \left(\frac{1}{\sigma\sqrt{2\pi}} \int_{\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt \right) + (1-\lambda) \left(\frac{1}{\tau\sqrt{2\pi}} \int_{\infty}^{x} e^{-\frac{(t-\nu)^2}{2\tau^2}} dt \right),
\end{aligned}$$

and the moment generating function is

$$M(t) = \sum_{i=0}^{\infty} \frac{m_i}{i!} t^i = \lambda e^{\mu t + \sigma^2 t^2/2} + (1-\lambda) e^{\nu t + \tau^2 t^2/2}.$$

Note that the m_i's are polynomial maps in the five parameters and $M(t)$ is defined on some interval $(-a, a)$. Thus M can be seen as a function $M\colon (-a,a) \times P \to \mathbb{R}$. The statement that these three functions characterize the distribution means that the equivalence relations they induce on $P = [0,1] \times \mathbb{R}^4_{\geq 0}$ coincide with the model-data equivalence relation. By the $2r+1$ result [52], it follows that for generic $x_1, \ldots, x_{11}$ and $t_1, \ldots, t_{11}$ each of the functions

$$\begin{aligned} \phi_1\colon P &\to \mathbb{R}^{11} \\ p &\mapsto (f(x_1,p), \ldots, f(x_{11},p)), \end{aligned}$$

$$\begin{aligned} \phi_2\colon P &\to \mathbb{R}^{11} \\ p &\mapsto (F(x_1,p), \ldots, F(x_{11},p)), \end{aligned}$$

and

$$\begin{aligned} \phi_3\colon P &\to \mathbb{R}^{11} \\ p &\mapsto (M(t_1,p), \ldots, M(t_{11},p)) \end{aligned}$$

also induce the model-data equivalence relation. Let $X_1, ..., X_K$ denote a random sample from the distribution. As the moment generating function can be estimated from the sample via $\frac{1}{K}\sum_{i=1}^{K} e^{tX_i}$, the map ϕ_3 is a model prediction map. As the cumulative distribution map can be estimated by the empirical distribution function, ϕ_2 is also a model prediction map. The probability density function can also, in principle, be indirectly estimated from the sample by numerically deriving the empirical distribution function that estimates the cumulative distribution function. Thus, ϕ_1 can also be considered as a model prediction map.

This model also has algebraic model prediction maps. Indeed, the set of moments $\{m_i \mid i \geq 0\}$ determines M, which implies that this set of polynomial functions $P \to \mathbb{R}$ will also induce the model-data equivalence relation. To obtain an algebraic model prediction map it will suffice to find a finite separating set $E \subset \mathbb{R}[m_i \mid i \geq 0] \subseteq \mathbb{R}[\lambda, \mu, \sigma, \nu, \tau]$, that is, a set E such that whenever two points of $\mathbb{R}^5$ are separated by some m_i, there is an element of E that separates them (see [31] for a treatment of separating sets for rings of functions). As $\mathbb{R}[\lambda, \mu, \sigma, \nu, \tau]$ is a finitely generated $\Bbbk$-algebra, by [31, Theorem 2.1] finite separating sets exist, and for d large enough the first $d+1$ moments $m_0, m_1, \ldots, m_d$ will form a separating set. In fact, through careful algebraic manipulations it is possible to show that the first 7 moments already form a separating set (see [2, Section 3] or [34]). As it is possible to estimate moments from data (via the sample moments $\frac{1}{K}\sum_{i=1}^{K} X_i^j$ for $j \geq 1$), we have a fourth model prediction map

$$\phi_4\colon [0,1] \times \mathbb{R}^4_{\geq 0} \to \mathbb{R}^6$$

$$p \mapsto (m_1, m_2, m_3, m_4, m_5, m_6).$$

◁

Let M be given by a real-analytic ODE system with time series data or an explicit time dependent model with time series data. Then by the $2r+1$ result [52], we know that there exist model prediction maps that capture all the time series information. For the explicit models it is simply a matter of choosing timepoints. For ODE systems, we would in principle need an exact solution. First, some examples of explicit time dependent models:

Example 3.10 (Fitting points to a line). By the discussion in Example 2.5, the model-data equivalence relation coincides with the equivalence relation induced by evaluating the variable x at the timepoints $t_1 = 0$ and $t_2 = 1$. As there is an invertible linear transformation taking any two distinct timepoints (t_1, t_2) to $(0, 1)$, any choice of two timepoints will give a model prediction map

$$\begin{aligned} \phi_{t_1,t_2}\colon \quad & \mathbb{R}^2 \to \mathbb{R}^2 \\ & (a_0, a_1) \mapsto (a_0 + t_1 a_1, a_0 + t_2 a_1). \end{aligned}$$

Each corresponding set of model predictions, that is the image of ϕ_{t_1,t_2}, actually fill up $\mathbb{R}^2$. The set of model predictions we would obtain by taking more timepoints would still be isomorphic to $\mathbb{R}^2$. ◁

Example 3.11 (Sum of exponentials). By the $2r+1$ result [52], any generic choice of 5 timepoints will provide a model prediction map, but as we saw in Example 2.11, two timepoints suffice. As in the paper [57], we use the three timepoints $t_1 = 1/3, t_2 = 1, t_3 = 3$ to define a model prediction map

$$\begin{aligned} \phi\colon \ & \mathbb{R}_{\geq 0} \to \mathbb{R}^3 \\ & (a, b) \mapsto (e^{-a/3} + e^{-b/3}, e^{-a} + e^{-b}, e^{-3a} + e^{-3b}). \end{aligned}$$

The image of ϕ, the corresponding set of model predictions, is a surface with a boundary given by the image of the line $\{(a, b) \mid a = b\}$. A set of model predictions obtained by measuring at two timepoints will consist of a closed subset of the positive quadrant of $\mathbb{R}^2$. ◁

For an ODE system with time series data, if we have an exact solution then we can easily construct a model prediction map as in the explicit time dependent case. In the absence of a solution, it may still be possible to construct a model prediction map, at least on a dense open subset of parameter space. For example, the coefficients of the input-output equations used in the differential algebra approach to obtain an exhaustive summary can be estimated from data (see for example [7, p. 17]). Hence, in this case one can construct a rational model manifold. For example, in Example 2.15 the map $\phi\colon U \to \mathbb{R}^5$, when seen as a rational map on the whole parameter space, is a rational model prediction map. In general, however, the best one can do is solve the ODE system

numerically and build a *numerical model prediction map* as is done in the sloppiness literature [56, 57, 58, 55]. A numerical model prediction map will provide some information on the model equivalence relation induced by an exact model prediction map; the quality of this information will depend on the quality of the numerics.

4. Sloppiness and its relationship to identifiability

We consider a model M with a fixed choice of model prediction map ϕ. A similar analysis can be made for a model with a generic model prediction map φ by replacing P with the domain of definition of φ where needed. For the rest of this paper we focus on models with model prediction maps.

We now consider the situation in which the data are model predictions corrupted by measurement noise with a known probability distribution. Hence, according to our assumption, the *noisy data* is the result of a random process. We define the *data space* $Z \subseteq \mathbb{R}^N$ to be the set of points of $\mathbb{R}^N$ that can be obtained as a corruption of the perfect data; how much it extends beyond the model predictions will depend on the support of the probability distribution of the measurement noise. The probability density function of the noisy data that can arise for the parameter value $p \in P$ is denoted by $\psi(p, \cdot)\colon Z \to \mathbb{R}$; it is the probability density of observing data $z \in Z$, which, for each $p \in P$, depends on the model prediction $\phi(p)$ rather than depending directly on the parameter p.

The *Kullback-Leibler divergence*, used in probability and information theory, quantifies the difference between two probability distributions [32]. We define a premetric on parameter space via the Kullback-Leibler divergence:

$$d(p, p') := \int_Z \psi(p, z) \log\left(\frac{\psi(p, z)}{\psi(p', z)}\right) dz. \tag{4}$$

Gibb's Inequality [16] proves that the Kullback-Leibler divergence is nonnegative, and zero only when the two probability distributions are equal on a set of probability one. It follows that d is a *premetric*, that is, $d(p, p') \geq 0$ and $d(p, p) = 0$. Furthermore, $d(p, p') = 0$ if and only if the probability distributions $\psi(p, \cdot)$ and $\psi(p', \cdot)$ are equal on a set of probability one, which is equivalent to $\phi(p) = \phi(p')$, since the dependence of ψ on p is only via the model prediction $\phi(p)$. Note that in general the Kullback-Leibler divergence and the premetric d are not symmetric and do not satisfy the triangle inequality.

Example 4.1 (The case of additive Gaussian measurement noise). Suppose the observations of a model prediction are distributed as follows:

$$z \sim \mathcal{N}(\phi(p), \Sigma), \tag{5}$$

where $\mathcal{N}(\phi(p), \Sigma)$ denotes a multivariate Gaussian distribution with mean $\phi(p) \in \mathbb{R}^N$ and covariance matrix Σ, a $N \times N$ positive semi-definite matrix. This is equivalent to specifying that $z = \phi(p) + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \Sigma)$, that is, the measurement noise is additive and Gaussian. We let K be the number of experimental replicates, or the size of the

sample. The density of a multivariate Gaussian then gives $\psi(p,\cdot)$ as

$$\psi(p,z) = (2\pi)^{-\frac{NK}{2}} |\Sigma|^{-\frac{K}{2}} \exp\left(-\frac{K}{2}\big\langle (z-\phi(p)), \Sigma^{-1}\big(z-\phi(p)\big)\big\rangle\right),$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product. The computation of (4) then yields :

$$d(p,p') = \frac{K}{2}\Big\langle \phi(p)-\phi(p'), \Sigma^{-1}\big(\phi(p)-\phi(p')\big)\Big\rangle, \tag{6}$$

(details provided in [18]). Thus $d(p,p')$ is a weighted sum of squares, and so it is symmetric and satisfies the triangle inequality, and hence d is a pseudometric. In particular, if Σ is the identity matrix, then d is induced by half of the square of the Euclidean distance in data space. The pseudometric d is a metric exactly when the model is globally identifiable, since then $d(p,p') = 0 \Leftrightarrow \phi(p) = \phi(p') \Leftrightarrow p = p'$. ◁

It is often possible to equip parameter space with a metric, a natural choice being the Euclidean metric inherited from the ambient $\mathbb{R}^r$. For instance our model might be of a chemical reaction network, where the coordinates of the parameter correspond to the positive, real-valued rate constants associated with particular chemical reactions. In this case, a reasonable choice of reference metric is the Euclidean distance between different points in the positive real quadrant. The reference metric on parameter space may not be Euclidean. For example, the natural metric on tree space that arises in Phylogenetics, the BHV metric, is non-Euclidean [6].

We can now offer a new precise, but qualitative, definition of sloppiness. We discuss two different quantifications in the following two sections:

Definition 4.2. Let (M,ϕ,ψ,d_P) be a mathematical model with a choice of model prediction map, a specific assumption on the probability distribution of the noisy data, and a choice of reference metric on P. We say that (M,ϕ,ψ,d_P) is *sloppy* at p_0 if in a neighborhood of p_0 the premetric d diverges significantly from the reference metric on parameter space.

4.1. Infinitesimal Sloppiness

We first provide the generally accepted and original quantification of sloppiness found in the literature, which we explain in terms of our new qualitative definition of sloppiness (see Definition 4.2). The sloppiness literature makes the implicit assumption that the reference metric on parameter space is the standard Euclidean metric, and we make the same assumption in this section.

Fix $p_0 \in P$ and consider the map $d(\cdot,p_0)\colon P \to \mathbb{R}_{\geq 0}$ mapping p to $d(p,p_0)$. Suppose that $d(\cdot,p_0)$ is twice continuously differentiable in a neighborhood of p_0. By definition, $d(p_0,p_0) = 0$, and furthermore p_0 is a local minimum of $d(\cdot,p_0)$, implying a null Jacobian. Therefore an approximation of $d(p,p_0)$ for p in a neighborhood of p_0 is given by the Taylor expansion

$$d(p,p_0) = \frac{1}{2}\Big\langle (p-p_0), \big(\nabla_p^2 d(p,p_0)\big)\,|_{p=p_0}(p-p_0)\Big\rangle + \mathcal{O}(\|(p-p_0)\|_2), \tag{7}$$

where $\|\cdot\|_2$ is the Euclidean norm and $\nabla_p^2 d(p,p_0)$ is the Hessian of the function $d(p,p_0)$, that is, the matrix $\left(\frac{\partial^2}{\partial p_i \partial p_j} d(p,p_0)\right)_{i,j}$ of second derivatives with respect to the coordinate parameters. This Hessian evaluated at $p=p_0$ is known as the *Fisher Information Matrix (FIM) at* p_0.

Local minimality of $d(\cdot,p_0)$ at p_0 ensures that the matrix $\left(\nabla_p^2 d(p,p_0)\right)|_{p=p_0}$ is positive semidefinite, and so the FIM at p_0 induces a pseudometric on parameter space

$$\begin{aligned} d_{\mathrm{FIM,p_0}} \colon P \times P &\to \mathbb{R}_{\geq 0} \\ (p,p') &\mapsto \frac{1}{2}\left\langle (p-p'), \left(\nabla_p^2 d(p,p_0)\right)|_{p=p_0}(p-p')\right\rangle. \end{aligned}$$

Note that the pseudo-metric $d(\cdot,p_0)$ is not the Fisher Information metric. When the FIM is positive definite, the Fisher Information metric is the Riemannian metric induced by the FIM by computing the line integral of the geodesic linking two parameters $p,p' \in P$ [1].

Example 4.3 (The case of additive Gaussian measurement noise). In the sloppiness literature, measurement noise is assumed Gaussian, as in Example 4.1, and for $K=1$ the FIM $\left(\nabla_p^2 d(p,p_0)\right)|_{p=p_0}$ is known as the *sloppiness matrix at* p_0. Explicitly, the sloppiness matrix is

$$\left(\nabla_p^2 d(p,p_0)\right)|_{p=p_0} = \frac{1}{2}\left((\nabla_p \phi(p))|_{p=p_0}\right)^T \Sigma^{-1} \left((\nabla_p \phi(p))|_{p=p_0}\right), \tag{8}$$

where $(\nabla_p \phi(p))|_{p=p_0}$ denotes the Jacobian of ϕ with respect to the coordinate parameters evaluated at $p=p_0$. ◁

Remark 4.4 (Structural identifiability and the FIM). The FIM is intimately linked to structural identifiability. Indeed, a result of Rothenberg [48, Theorem 1] shows that M is locally identifiable if and only if the FIM is full rank at some p_0. If we assume additive Gaussian noise, then Equation 8 implies that the rank r_0 of the FIM at p_0 is equal to the rank of the Jacobian of ϕ at p_0, and so for generic p_0, the dimension of the connected component of p_0 in its equivalence class is $r-r_0$ (cf discussion near [15, Equation 85]). As one can compute the rank of the FIM by computing the singular value decomposition and employing a sound threshold [22], the FIM can then be used to numerically determine the dimension of generic equivalence classes. Further approaches for giving probabilistic, and sometimes guaranteed bounds on identifiability using symbolic computation at specific parameters have been developed and applied in [3, 49, 30].

The Taylor expansion (7) shows that, for parameters very near p_0, the premetric d is approximately given by the pseudometric $d_{\mathrm{FIM,p_0}}$. Therefore, in a neighborhood of p_0, the map ϕ giving the model predictions is maximally sensitive to infinitesimal perturbations in the direction of the eigenvector of the maximal eigenvalue of the FIM at p_0, referred to as the *stiffest* direction at p_0. The direction of the eigenvector of the minimal eigenvalue of the FIM at p_0, which gives the perturbation direction to which ϕ is minimally sensitive, is known as the *sloppiest* direction at p_0.

Definition 4.5. Let (M, ϕ, ψ, d_2) be a mathematical model with a choice of model prediction map, a specific assumption on measurement noise, and the Euclidean metric as a reference metric on P. We say that (M, ϕ, ψ, d_2) is *infinitesimally sloppy* at a parameter p_0 if there are several orders of magnitude between the largest and smallest eigenvalues of the FIM at p_0. We define the *infinitesimal sloppiness at* p_0 to be the condition number of the FIM at p_0, that is, the ratio between its largest and smallest eigenvalues.

Remark 4.6. First note that this definition is only meaningful when the FIM at p_0 is full rank. In this case, the condition number of the FIM at p_0 corresponds to the aspect ratio of the level curves of $d_{\mathrm{FIM,p_0}}$, which is one way to quantify how far these level curves are from Euclidean spheres. Thus, using the condition number of the FIM as a quantification of sloppiness implies that the reference metric on P is the Euclidean metric.

The FIM possesses attractive statistical properties. Suppose (M, ϕ, ψ, d_2) is locally identifiable and that maximum likelihood estimates exist generically, that is, for almost all z, there are parameters minimizing the negative log-likelihood: $\hat{p}(z) = \min_{p \in P}(-\log \psi(p, z))$. Let $z \in Z$ be a generic data point and let $\hat{p}(z)$ be the unique maximum likelihood estimate. Suppose that the "true" parameter is p_0, that is, z is a corruption of the model prediction $\phi(p_0)$. When the FIM at p_0 is invertible, the Cramer-Rao inequality [11, Section 7.3] implies that

$$[\left(\nabla_p^2 d(p, p_0)\right)|_{p=p_0}]^{-1} \preceq \mathrm{Cov}_{p_0}\, \hat{p}(z), \tag{9}$$

where

$$\mathrm{Cov}_{p_0}\, \hat{p}(z) := \int_Z \hat{p}(z)\hat{p}^T(z)\psi(p_0, z)\ dz - \left(\int_Z \hat{p}(z)\psi(p_0, z)\ dz\right)\left(\int_Z \hat{p}(z)\psi(p_0, z)\ dz\right)^T$$

is the covariance of the maximum likelihood estimate with respect to measurement noise, and $A \preceq B$ if and only if $B - A$ is positive semi-definite. This inequality provides an explicit link between the uncertainty associated with parameter estimation and the geometry of the negative log likelihood. Meanwhile, the sensitivity of ϕ is related to the uncertainty associated with parameter estimation via (9). The asymptotic normality of the maximum likelihood estimates implies that the Cramer-Rao inequality (9) tends to *equality* as K tends to infinity [11, Section 10.7]. Formally,

$$\lim_{K \to \infty} [\left(\nabla_p^2 d(p, p_0)\right)|_{p=p_0}]^{-1} = \mathrm{Cov}_{p_0}\, \hat{p}(z). \tag{10}$$

The list of regularity conditions required for (9) and (10) to hold are provided in [11, Section 7.3], and are easily satisfied in practice.

Remark 4.7. A sufficient condition for $d_{\mathrm{FIM,p_0}}$ to be a good approximation for the premetric d on a neighborhood of p_0 is to have a very large number of replicates. In practice, however, questions of cost and time mean that the number of replicates is often very small. Accordingly, the sloppiness literature generally assumes the number of experiments is one ($K = 1$), though the effect of increasing experimental replicates in mitigating sloppiness has been explored in [4].

Example 4.8 (Fitting points to a line). We revisit once more the model first considered in Example 2.5. We consider the model prediction map obtained by evaluating at timepoints $t_1 = 0$ and $t_2 = 1$ as in Example 3.10. We assume that we are in the presence of additive Gaussian error with covariance matrix $\Sigma = I_2$ equal to the identity matrix as discussed in Examples 4.1 and 4.3. As in Example 4.1 the premetric d is induced by half the square of the Euclidean distance on the data space $\mathbb{R}^2$. We can explicitly determine d:

$$\begin{aligned} d((a_0, a_1), (a_0', a_1')) &= \frac{1}{2}\left((a_0 - a_0')^2 + (a_0 + a_1 - a_0' - a_1')^2\right) \\ &= \frac{1}{2}\left(2(a_0 - a_0')^2 + 2(a_0 - a_0')(a_1 - a_1') + (a_1 - a_1')^2\right) \\ &= \frac{1}{2}\left\langle \begin{pmatrix} a_0 - a_0' \\ a_1 - a_1' \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} a_0 - a_0' \\ a_1 - a_1' \end{pmatrix} \right\rangle. \end{aligned}$$

We see that d itself is a weighted sum of squares given by a positive definite matrix, and so d is a metric. As the positive definite matrix giving this sum of squares is constant throughout parameter space, it follows that the sloppiness of the model is also constant throughout parameter space. Note that the same phenomenon would happen for any model such that the model manifold is given by an injective linear map (see Proposition 4.9 below). In particular, the same situation would arise when considering the problem of fitting points to any polynomial curve, as the corresponding model prediction map will be linear.

We next compute the FIM. The map $d(\cdot, (b_0, b_1))$ is given by

$$d((a_0, a_1), (b_0, b_1)) = \frac{1}{2}\left((b_0 - a_0)^2 + (b_0 + b_1 - a_0 - a_1)^2\right)$$

and so its Hessian, that is, the FIM is

$$\frac{1}{2}\begin{pmatrix} \frac{\partial^2}{\partial a_0^2} d((a_0, a_1), (b_0, b_1)) & \frac{\partial^2}{\partial a_0 \partial a_1} d((a_0, a_1), (b_0, b_1)) \\ \frac{\partial^2}{\partial a_1 \partial a_0} d((a_0, a_1), (b_0, b_1)) & \frac{\partial^2}{\partial a_0^2} d((a_0, a_1), (b_0, b_1)) \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$$

We conclude that in this case, the pseudometric $d_{\mathrm{FIM},(a_0,a_1)}$ coincides with d on the entire parameter space, which we will see in Proposition 4.9 is a consequence of the linearity of the model prediction map. ◁

Proposition 4.9. Let $(M, \phi, \mathcal{N}(\phi(p), \Sigma), d_2)$ be a mathematical model with parameter space $P \subseteq \mathbb{R}^r$, a choice of model prediction map, additive Gaussian noise with covariance matrix Σ, and the Euclidean metric as a reference metric. If the model prediction map $\phi\colon P \to \mathbb{R}^N$ is linear, then $d_{\mathrm{FIM},p_0} = d$ for all $p_0 \in P$.

Proof. Our assumption that ϕ is linear implies that there is a $N \times r$ matrix A with real entries such that $\phi(p) = Ap$. By the discussion in Example 4.1, we have

$$d(p', p) = \frac{K}{2}\left\langle (Ap' - Ap), \Sigma^{-1}(Ap' - Ap) \right\rangle$$

$$= \frac{K}{2}\Big\langle (A(p'-p)), \Sigma^{-1}(A(p'-p))\Big\rangle$$
$$= \frac{K}{2}\Big\langle (p'-p), (A^T\Sigma^{-1}A)(p'-p)\Big\rangle.$$

On the other hand the FIM is given by

$$\begin{aligned}\nabla_p^2 d(p,p_0) &= \nabla_p^2\left(-\frac{NK}{2}\log(2\pi)+\frac{K}{2}\log(|\Sigma|)+\frac{K}{2}\Big\langle (Ap_0-Ap),\Sigma(Ap_0-Ap)\Big\rangle\right)\\ &= \frac{K}{2}\nabla_p^2\left(\Big\langle (Ap_0-Ap),\Sigma^{-1}(Ap_0-Ap)\Big\rangle\right)\\ &= \frac{K}{2}\nabla_p^2\left(\Big\langle (Ap),\Sigma^{-1}(Ap)\Big\rangle\right)\\ &= \frac{K}{2}A^T\Sigma^{-1}A,\end{aligned}$$

completing the proof.

Example 4.10 (Linear parameter-varying model). We consider a standard model arising in control theory, which falls under the case of real analytic time dependent models. Specifically, we consider models of the form

$$\begin{aligned}\dot{x} &= A(p)x,\\ y &= Cx,\\ x(0) &= x_0,\end{aligned}$$

where $A(p)$ is a $m\times m$ matrix with polynomial dependence on the parameter $p\in\mathbb{R}^{r-1}$, C is a known fixed $n\times m$ matrix with real coefficients, and y is the measurable output. Note that y depends on the initial condition x_0, which we will consider as an extension of parameter space. We assume further that $A(p)$ is Hurwitz for all p considered. We denote by $y(t,(p,x_0))$ the output of the system at time t, given the parameter (p,x_0). If we measured the system at a finite number of time-points, assuming Gaussian noise-corruption, then the distance function $d((p,x_0),(p',x_0'))$ would be the Euclidean distance between the model predictions at the chosen set of timepoints.

For any pair of parameters (p,x_0) and (p',x_0'), the following integral can be explicitly computed and is a rational function of (p,x_0) and (p',x_0') (cf [43, Theorem 1], which assumes that $A(p)$ is linear in the parameters, but whose proof holds more generally):

$$d_\infty((p,x_0),(p',x_0')) := \int_0^\infty \|y(t,(p,x_0))-y(t,(p',x_0'))\|_2^2\ dt.$$

Note that $d_\infty((p,x_0),(p',x_0'))$ is equal to the L^2 norm of the function $y(t,(p,x_0))-y(t,(p',x_0'))$, and so $d_\infty((p,x_0),(p',x_0'))=0$ if and only if $y(t,(p,x_0))=y(t,(p',x_0')$ for almost all t. As y is real-analytic, it then follows that $y(t,(p,x_0))=y(t,(p',x_0')$ for all t. Therefore $d_\infty((p,x_0),(p',x_0'))=0$ if and only if $(p',x_0')\sim_{M,z}(p,x_0)$, and so the equivalence class of (p,x_0) is given by the zeros (p',x_0') of the rational function $d_\infty((p,x_0),(p',x_0'))$. ◁

4.2. Multiscale sloppiness

We now present a quantification of sloppiness that holds for non-Euclidean reference metric and is better suited to the presence of noninfinitesimal noise. In this section, we sometimes make the assumption that for generic $p_0 \in P$, there is a neighborhood of p_0 where the reference metric d_P is strongly equivalent to the Euclidean metric inherited by P as a subset of $\mathbb{R}^r$. The BHV metric [6] mentioned at the beginning of the section satisfies this property.

In Section 4.1 we saw how the FIM approximates the premetric d in the limit of decreasing magnitude of parameter perturbation, which is realizable in the limit of increasing experimental replicates or sample size. In a practical context, however the limit of increasing replicates may not be valid. Indeed, examples are provided in [26] and [29] of models for which the uncertainty of parameter estimation is poorly approximated by the FIM. Even when the approximation is valid, numerical errors in sloppiness quantification are often significant, due to the ill-conditioning of the FIM [60]. We describe a second approach called *multiscale sloppiness* introduced in [44] for models given by ODE systems with time series data under the assumption of additive Gaussian noise and with the standard Euclidean metric as a reference metric. We extend this quantification of sloppiness to a more general setting.

Definition 4.11 (Multiscale Sloppiness)**.** Consider a model (M, ϕ, ψ, d_P) with a choice of model prediction map ϕ, a specific assumption on measurement noise, and a choice of reference metric on P. We define the *δ-sloppiness at p_0* to be

$$\mathcal{S}_{p_0}(\delta) := \frac{\sup_{p\in P}\{d(p,p_0) \mid d_P(p,p_0) = \delta\}}{\inf_{p\in P}\{d(p,p_0) \mid d_P(p,p_0) = \delta\}}$$

If d_P is strongly equivalent to the Euclidean metric on a neighborhood of p_0, then for δ sufficiently small, the (non-unique) *maximally* and *minimally* disruptive parameters at length scale δ at the point $p_0 \in P$ are the elements of the sets

$$D^{max}_{p_0}(\delta) = \arg\max_{p\in P} d(p,p_0) : d_P(p,p_0) = \delta \tag{11}$$

$$D^{min}_{p_0}(\delta) = \arg\min_{p\in P} d(p,p_0) : d_P(p,p_0) = \delta, \tag{12}$$

respectively. In this case, the δ-sloppiness at p_0 is

$$\mathcal{S}_{p_0}(\delta) = \frac{d(p^{max}_{p_0}(\delta), p_0)}{d(p^{min}_{p_0}(\delta), p_0)}, \tag{13}$$

where $p^{max}_{p_0}(\delta) \in D^{max}_{p_0}(\delta)$ and $d(p^{min}_{p_0}(\delta) \in D^{min}_{p_0}(\delta)$.

Note that since the set $\{d(p,p_0) \mid p \in P \text{ and } d_P(p,p_0) = \delta\}$ is a closed set of real numbers with a lower bound (zero), the infimum is actually a minimum, hence $D^{min}_{p_0}(\delta)$ is always well-defined.

Remark 4.12. Computation of δ-sloppiness would seem to require the solution of a (possibly nonlinear, nonconvex) optimization program for each $\delta > 0$. However, assuming the reference metric on parameter space is the Euclidean distance and that we are in the presence of additive Gaussian noise, finding $p_{p_0}^{min}(\delta) \in D_{p_0}^{min}(\delta)$ for continuous ranges of δ can be formulated as the solution of an optimal control problem relying on solving a Hamiltonian $dH/dp = 0$ as described in [44, Section 5]. With this method, computation of δ-sloppiness is possible for large, nonlinear systems of ODE. Note that this formulation as an optimal control problem does not fundamentally rely on the assumption of a Euclidean metric on parameter space, and so the principle likely applies to more general classes of metric.

If we choose the usual Euclidean distance as the reference metric on parameter space, then as the length-scale δ goes to zero, infinitesimal and multiscale sloppiness coincide:

$$\lim_{\delta \to 0} \mathcal{S}_{p_0}(\delta) = \frac{\lambda^{max}\left(\nabla_p^2 d(p, p_0)\right)|_{p=p_0}}{\lambda^{min}\left(\nabla_p^2 d(p, p_0)\right)|_{p=p_0}},$$

where λ^{max} and λ^{min} denote the maximal and minimal eigenvalues of their argument. Indeed, the Taylor expansion (7) implies that as p approaches p_0, $d(p, p_0)$ approaches $d_{\mathrm{FIM},p_0}(p, p_0)$, and so the level sets $\{p \in P \mid d(p, p_0) = \delta\}$ tend to the level sets $\{p \in P \mid d_{\mathrm{FIM},p_0}(p, p_0) = \delta\}$ as δ goes to zero.

Multiscale sloppiness, or more precisely the denominator of $\mathcal{S}_{p_0}(\delta)$, is closely related to structural identifiability:

Theorem 4.13. Let (M, ϕ, ψ, d_P) be a mathematical model with a choice of model prediction map, a specific assumption on measurement noise, and a choice of reference metric d_P, which we assume is strongly equivalent to the Euclidean metric. The equivalence class $[p_0]_{\sim_{M,\phi}}$ of the parameter p_0 has size one if and only if $\inf_{p \in P}\{d(p, p_0) \mid d_P(p, p_0) = \delta\} > 0$ for all $\delta > 0$.

Proof. Suppose that the equivalence class $[p_0]$ of p_0 has size one, then for any other parameter p, we will have $d(p, p_0) > 0$. In particular, this will hold for $p_{p_0}^{min}(\delta) \in D_{p_0}^{min}(\delta)$, for any $\delta > 0$. Hence $\inf_{p \in P}\{d(p, p_0) \mid d_P(p, p_0) = \delta\} = d(p_{p_0}^{min}(\delta), p_0) > 0$.

Suppose on the other hand that $\inf_{p \in P}\{d(p, p_0) \mid d_P(p, p_0) = \delta\} > 0$ for all $\delta > 0$ and suppose, for a contradiction that $p \in [p_0]$ is distinct from p_0. Set $\delta' := d_P(p, p_0)$. As $p \neq p_0$, we have $\delta' > 0$ and

$$0 = d(p, p_0) \geq \inf_{p \in P}\{d(p, p_0) \mid d_P(p, p_0) = \delta'\},$$

which is a contradiction, since $\inf_{p \in P}\{d(p, p_0) \mid d_P(p, p_0) = \delta'\} > 0$.

Example 4.14 (Sum of exponentials). We highlight that sloppiness is a local property: it depends on the point in parameter space and the precise choice of timepoints. In this spirit, let us revisit Example 2.6, again adding Gaussian measurement noise with identity covariance and taking the model prediction map to be evaluating at timepoints

$\{1/3, 1, 3\}$. We are in the situation considered in Example 4.1 and so d is again half the squared Euclidean distance between model predictions:

$$d\left((a,b),(a',b') \right) = \frac{1}{2}\|\phi(a,b) - \phi(a',b')\|_2^2,$$

where

$$\phi(a,b) = (e^{-a/3} + e^{-b/3}, e^{-a} + e^{-b}, e^{-3a} + e^{-3b}).$$

The Jacobian of the model prediction map at (a_0, b_0) is therefore given as

$$(\nabla_{a,b}\phi(a,b))|_{(a,b)=(a_0,b_0)} = \begin{pmatrix} -\frac{1}{3}e^{-a_0/3} & -e^{-a_0} & -3e^{-3a_0} \\ -\frac{1}{3}e^{-b_0/3} & -e^{-b_0} & -3e^{-3b_0} \end{pmatrix}$$

The FIM at (a_0, b_0), in this case, will be given as

$$\begin{pmatrix} \frac{1}{9}e^{-2a_0/3} + e^{-2a_0} + 9e^{-6a_0} & \frac{1}{9}e^{-(a_0+b_0)/3} + e^{-(a_0+b_0)} + 9e^{-3(a_0+b_0)} \\ \frac{1}{9}e^{-(a_0+b_0)/3} + e^{-(a_0+b_0)} + 9e^{-3(a_0+b_0)} & \frac{1}{9}e^{-2b_0/3} + e^{-2b_0} + 9e^{-6b_0} \end{pmatrix}$$

We compute infinitesimal sloppiness and δ-sloppiness of (M, ϕ) at $p_0 = (a, b) = (4, 1/8)$ in Figure 4: the difference between these notions of sloppiness becomes clear.

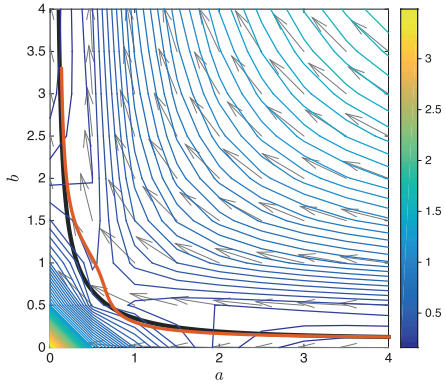


Figure 4: Infinitesimal sloppiness vs δ-sloppiness of sum of exponential (Example 4.14). For a given parameter $p_0 = (a = 4, b = 1/8)$ and time points $\{1/3, 1, 3\}$, level sets of d are drawn (colors). The vector field consisting of the eigenvector corresponding to the largest eigenvalue of the FIM is plotted across the grid. We compare the flow of this vector field initialized at p_0 (gray curve), with the most delta-sloppy parameters with respect to p_0 over a range of δ (orange curve).

Figure 5 illustrates how the change of model prediction map, in this case different choices of timepoints, changes the premetric d. This suggests that sloppiness should be taken into consideration when designing an experiment: some choices of timepoints will allow for better quality parameter estimation. Figure 6, on the other hand, illustrates how the premetric d changes in parameter space. In particular, these two figures illustrate that unlike identifiability, sloppiness is not a global property of a model. ◁

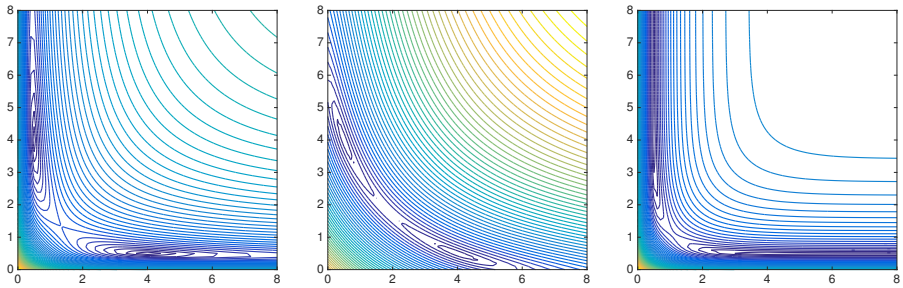


Figure 5: Sloppiness for different choices of model prediction map for the sum of exponentials (Example 4.14). For a given parameter $p_0 = (4, 1/2)$, we draw the level curves of $\sqrt{d(\cdot, (4, 1/2))}$ for timepoints $\{1/3, 1, 3\}$ on the left, for timepoints $\{1/9, 1/3\}$ in the center, and for timepoints $\{1, 3\}$ on the right are shown. Taking the square root changes the spacing of the level curves, but not on their shape.

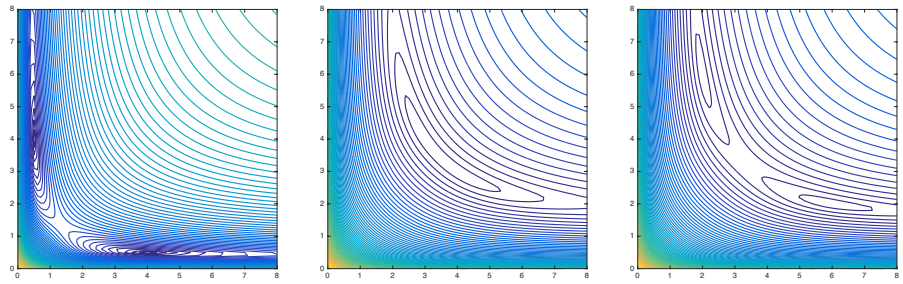


Figure 6: Sloppiness at different parameters given a choice of model prediction map for the sum of exponentials (Example 4.14). With the model prediction map given by timepoints $\{1/3, 1, 3\}$, we draw the level curves of $\sqrt{d(\cdot, (4, 1/2))}$ on the left, of $\sqrt{d(\cdot, (3, 3))}$ in the center, and of $\sqrt{d(\cdot, (6, 2))}$on the right are shown. Taking the square root changes the spacing of the level curves, but not on their shape.

4.3. Sloppiness and practical identifiability

Determining the practical identifiability of a model corresponds to asking whether one can arrive to some estimate of the parameter from noisy data, that is, whether based on an assumption on measurement noise, noisy data constrains the parameter value to a bounded region of parameter space. Part of the literature uses the FIM in the manner of infinitesimal sloppiness to define practical identifiability(see for example [59, 13]), but we will see in Example 4.19 that this method of evaluating practical identifiability can lead to problems. We thus favor an approach more in line with Raue et al [46].

Practical identifiability depends on the method used for parameter estimation. We focus on practical identifiability for maximum likelihood estimation, one of the most widely used methods for parameter estimation (see, for example [35]). Accordingly, in the remaining of this section, we consider models (M, ϕ, ψ, d_P) with a choice of model prediction map, a specific assumption of the probability distribution of measurement noise and a choice of

reference metric on P such that maximum likelihood estimates exist for generic data.

For the noisy data point $z_0 \in Z$, supposing the existence of a unique maximum likelihood estimate $\hat{p}(z_0)$ (i.e. supposing the model is generically identifiable, see Proposition 4.15 below), we define an ϵ-*confidence region* $U_\epsilon(z_0)$ as follows:

$$U_\epsilon(z_0) = \{p \in P \mid -\log \psi(p, z_0) < \epsilon\}.$$

The ϵ-confidence region therefore denotes the set of parameters that fit the data at least as well as some cutoff quality of fit, predicated on ϵ. The set $U_\epsilon(z_0)$ is often known as a Likelihood-based confidence region [59, 11], and is intimately connected with the Likelihood Ratio Test: Suppose we had a null hypothesis $\mathcal{H}_0$ that data z_0 was generated (modulo noise) through a parameter p_0, and we wished to test the alternative hypothesis $\mathcal{H}_1$ that z_0 was generated through some other parameter. By definition, a Likelihood Ratio test would reject the null hypothesis when

$$\Lambda(p_0, z_0) := \frac{\psi(p_0, z_0)}{\psi(\hat{p}(z_0), z_0)} \leq k^*,$$

where k^* is a critical value, with the significance level α equal to the probability $\Pr(\Lambda(z_0) \leq k^* | \mathcal{H}_0)$ of rejecting the null hypothesis when it is in fact true. The set of parameters such that the nul hypothesis is not rejected at significance level α is

$$\{p' \in P \mid log\psi(p', z_0) < -\log -\psi(\hat{p}(z_0), z_0) - \log k^*\},$$

that is, $U_\epsilon(z_0)$, where $\epsilon = -\log -\psi(\hat{p}(z_0), z_0) - \log k^*$.

Proposition 4.15 (closely related to [12, Theorem 2])**.** Let (M, ϕ, ψ) be a mathematical model with a model prediction map, and a specific assumption on measurement noise. Suppose that maximum likelihood estimates exist for generic data. If ϕ and ψ are real-analytic, then for almost all $z_0 \in Z$, the set of maximum likelihood estimates $\hat{p}(z)$, consists of exactly one equivalence class of $\sim_{M,\phi}$.

Proof. Let $z_0 \in Z$ be a generic data point. Solving the likelihood equation corresponds to finding the model prediction "closest" to the noisy data, as measured via the negative log-likelihood. We can assume without loss of generality that there is a unique solution to the likelihood equations. Indeed, under our assumptions, the set of data points where the closest model prediction is not unique will be contained in the zero set of analytic functions. Thus, the set of maximum likelihood estimates will consist of a single equivalence class. We can further assume that this equivalence class has generic size.

Remark 4.16. The ML degree [27], where the acronym "ML" stand for maximum likelihood, is defined as the number of complex solutions to the likelihood equations (for generic data). The ML degree is an upper bound for the number of solutions for the maximum likelihood equation, in particular it is an upper bound on the size of the equivalence classes when maximum likelihood estimates exist.

Even if for generically identifiable models the maximum likelihood estimate is unique with probability one, the parameter may not be identifiable in practice, meaning that noisy data does not constrain the parameter value to a bounded region of parameter space for a significant portion of the data space. More precisely, we refine the definition of Raue et al [46]:

Definition 4.17 (Practical identifiability)**.** Let (M, ϕ, ψ, d_P) be a mathematical model with a model prediction map, a specific assumption on measurement noise and a choice of reference metric d_P on P. Suppose that maximum likelihood estimates exist for generic data. Then (M, ϕ, ψ, d_P) is *practically identifiable at significance level* α if and only if for generic $z_0 \in Z$, there is a unique maximum likelihood estimate and the confidence region $U_\epsilon(z_0)$ is bounded with respect to the reference metric d_P, where ϵ satisfies

$$p' \in U_\epsilon(z_0) \Leftrightarrow \Pr\Big(-\log \psi(p', \hat{z}) < \epsilon \;\Big|\; \hat{z} \in Z \text{ is a corruption of } \phi(\hat{p}(z_0)\Big) = 1 - \alpha.$$

The model M is *practically unidentifiable at significance level* α if and only if there is a positive measure subset $Z' \subset Z$ such that for $z_0 \in Z'$, the confidence interval $U_\epsilon(z_0)$ is unbounded with respect to the reference metric d_P on P.

A model is practically identifiable at significance level α if generic data imposes that the parameter estimate belongs to a bounded region of parameter space, but this confidence region could be very large. Hence practical identifiability in this sense may not necessarily be completely satisfactory to the practitioner. One can further quantify practical identifiability to take into account the size of confidence regions, see for example [44].

Sloppiness and practical identifiability are complementary concepts. Practically identifiable models can be very sloppy, for example if the estimation of one component of the parameter is much more precise than that of another, see example below.

Example 4.18 (Practically identifiable, but sloppy)*.* Models with linear model prediction maps, Euclidean parameter space and standard additive Gaussian noise are always practically identifiable according to our definition, but these models can be arbitrarily sloppy.

We consider a model with 2-dimensional Euclidean parameter space and a linear model prediction map ϕ given by $(a, b) \mapsto (10^N a, b)$. We assume further that the measurement noise is Gaussian with identity covariance matrix. By Proposition 4.9, $d = d_{FIM,p_0}$ for any p_0 and at any scale, the level curves of d are ellipses with aspect ratio 10^N.

Our assumption of additive Gaussian noise implies that for any z_0, for each $\epsilon > 0$, the confidence interval $U_\epsilon(z_0)$ is an oval whose boundary ellipse is the level set of d centered at the maximum likelihood estimate $\hat{p}(z_0)$. Thus the confidence intervals $U_\epsilon(z_0)$ are bounded for any $\epsilon > 0$, and so the model is practically identifiable. ◁

In the following, we give an example of a model that is almost everywhere not sloppy at the infinitesimal scale, but is not practically identifiable. This model, however, exhibits some sloppiness at the non-infinitesimal scale. We see that the boundedness of level curves of d almost every where does not imply the boundedness of confidence intervals almost everywhere.

Example 4.19 (Not sloppy at the infinitesimal scale, but not practically identifiable). Consider the mathematical model $(M, \phi, \mathcal{N}(\phi(p), I_2), d_2)$ given by

$$\begin{aligned}\phi :&[1/2, \infty) \times \mathbb{R} \to \mathbb{R}^2 \\ &(a, b) \qquad\quad \mapsto \left(\frac{a}{a^2 + b^2}, -\frac{b}{a^2 + b^2} \right),\end{aligned}$$

additive Gaussian noise with identity covariance matrix, and parameter space $P = [1/2, \infty) \times \mathbb{R}$ equipped with the usual Euclidean metric.

The model prediction map ϕ is a conformal mapping that maps the closed half plane $[1/2, \infty) \times \mathbb{R}$ to the closed disc of radius 1 centered at $(1, 0)$ minus the origin. Since it is a conformal mapping, it preserves angles, and so infinitesimal circles are sent to infinitesimal circles. Under our assumptions on measurement noise and with the standard Euclidean metric as a reference metric on P, the model is not sloppy at all at the infinitesimal scale at parameters belonging to the open half plane $(1/2, \infty) \times \mathbb{R}$, but becomes increasingly sloppy at larger and larger scale, especially away from the parameter $(1/2, 0)$. The injectivity of the map ϕ on $P = [1/2, \infty) \times \mathbb{R}$ implies that the model (M, ϕ) is globally identifiable.

Our assumption on measurement noise implies that the maximum likelihood estimate is the parameter whose image is closest to the data point z_0, it will exist for any data point outside the closed half line $(-\infty, 0] \times \{0\}$. The confidence region $U_\epsilon(z_0)$ is then the preimage of the Euclidean open disc of radius ϵ centered at z_0. Whenever the closure of this open disc contains the origin, the corresponding confidence region will be unbounded. ◁

The final example illustrates that the uniqueness of the maximum likelihood estimate is independent of the boundedness of the confidence regions:

Example 4.20 (Bounded confidence regions but not practically identifiable). Consider the mathematical model $(M, \phi, \mathcal{N}(\phi(p), I_2), d_2)$ with model prediction map

$$\begin{aligned}\phi :&[1/2, \infty) \times \mathbb{R} \to \mathbb{R} \\ &(a, b) \qquad\quad \mapsto a^2 + b^2,\end{aligned}$$

additive Gaussian noise with identity covariance matrix, and parameter space $P = [1/2, \infty) \times \mathbb{R}$ equipped with the usual Euclidean metric. The equivalence class of the model-data equivalence relation are the concentric circles $\{(a, b) \mid a^2 + b^2 = r\}$ for $r \geq 0$, and so the model is generically non-identifiable. By Proposition 4.15, the set of maximum likelihood estimates for a generic $(a_0, b_0) \in \mathbb{R}$ is also a circle centered at the origin and the model is practically non-identifiable on any open neighborhood of (a_0, b_0). On the other hand, as the measurement noise is assumed to be Gaussian and the equivalence classes of $\sim_{M,\phi}$ are bounded, any confidence region will be bounded as well. Indeed, the confidence region $U_\epsilon((a_0, b_0))$ will be either an open disk $\{(a, b) \in \mathbb{R}^2 \mid a^2 + b^2 < a_0^1 + b_0^2 + \epsilon\}$, when $\epsilon > a_0^1 + b_0^2$, or an open ring $\{(a, b) \in \mathbb{R}^2 \mid a_0^1 + b_0^2 - \epsilon < a^2 + b^2 < a_0^1 + b_0^2 + \epsilon\}$, otherwise. ◁

5. Future of sloppiness

There are a number of interesting future directions for the theory and application of sloppiness. While we explained sloppiness via identifiability, this is only the beginning. An important next step is understanding sloppiness in the context of existing inference and uncertainty quantification theory. In terms of applications, there are some models where the reference metric on parameter space is non-Euclidean and we believe the computation of multiscale sloppiness can be adapted. While beyond the expertise of the authors, we would be excited to learn how the presented geometry of sloppiness extends to stochastic differential equations.

We highlighted how sloppiness is a local property, dependent on the parameter and timepoints of experiment. This dependence is reflected in model selection studies where a different model is selected depending on the choice of timepoints [51] or experimental stimulus dose [23]. We believe quantifying the shape of δ-sloppiness in relation to identifiability will have direct impact on parameter estimation.

In the last few years, researchers have successfully used FIM-based sloppiness to perform dimension reduction [58, 55, 21]. This is similar to the profile likelihood approach to dimension reduction [9], whose connection with identifiability was subsequently made [8]. This motivates an alternative understanding of when a model should be considered sloppy, namely when such model reduction is possible. Considerable more work is required in order to formalize this approach. A first step towards a definition is found in [33], where they propose predictive sloppiness. Predictive sloppiness is meant to be reparametrization invariant. However, how to obtain a closed form "exact" reparametrization remains an open problem.

6. Acknowledgements

HAH gratefully acknowledges the late Jaroslav Stark for posing this problem. The authors thank Carlos Améndola, Murad Banaji, Mariano Beguerisse Díaz, Sam Cohen, Ian Dryden, Paul Kirk, Terry Lyons, Chris Meyers, Jim Sethna, Eduardo Sontag, Bernd Sturmfels, and Jared Tanner for fruitful discussions. Additionally, we thank the anonymous referees for their helpful comments. This paper arises from research done while ED was a postdoctoral research assistant at the Mathematical Institute in Oxford funded by the John Fell Oxford University Press (OUP) Research Fund, and DVR was supported by the EPSRC Systems Biology Doctoral Training Center. ED is now supported by an Anne McLaren Fellowship from the University of Nottingham. HAH and DVR began discussions at the 2014 Workshop on Model Identification funded by KAUST KUK-C1-013-04. HAH was supported by EPSRC Fellowship EP/K041096/1 and now a Royal Society University Research Fellowship.

References

[1] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI; Oxford University Press, Oxford, 2000. Translated from the 1993 Japanese original by Daishi Harada.

[2] Carlos Améndola, Jean-Charles Faugère, and Bernd Sturmfels. Moment Varieties of Gaussian Mixtures. *J. Algebr. Stat.*, 7(1):14–28, 2016.

[3] Milena Anguelova. *Observability and identifiability of nonlinear systems with applications in biology*. PhD thesis, Chalmers University of Technology, 2007.

[4] Joshua F. Apgar, David K. Witmer, Forest M. White, and Bruce Tidor. Sloppy models, parameter uncertainty, and the role of experimental design. *Molecular Biosystems*, 6(10):1890–1900, oct 2010.

[5] Giuseppina Bellu, Maria Pia Saccomani, Stefania Audoly, and Leontina D'Angiò. Daisy: A new software tool to test global identifiability of biological and physiological systems. *Computer Methods and Programs in Biomedicine*, 88(1):52–61, 2007.

[6] Louis J. Billera, Susan P. Holmes, and Karen Vogtmann. Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, 27(4):733–767, 2001.

[7] François Boulier. Differential elimination and biological modelling. In *Gröbner bases in symbolic analysis*, volume 2 of *Radon Series on Computational and Applied Mathematics*, pages 109–137. Walter de Gruyter, Berlin, 2007.

[8] Andrew F Brouwer and Marisa C Eisenberg. The underlying connections between identifiability, active subspaces, and parameter space dimension reduction. *arXiv preprint arXiv:1802.05641*, 2018.

[9] Andrew F Brouwer, Rafael Meza, and Marisa C Eisenberg. Parameter estimation for multistage clonal expansion models from cancer incidence data: A practical identifiability analysis. *PLoS computational biology*, 13(3):e1005431, 2017.

[10] Kevin S. Brown and James P. Sethna. Statistical mechanical approaches to models with many poorly known parameters. *Physical Review E*, 68(2):021904, 2003.

[11] George Casella and Roger L. Berger. *Statistical Inference*. Duxbury, 2nd edition, 2002.

[12] E. A. Catchpole and B. J. T. Morgan. Detecting parameter redundancy. *Biometrika*, 84(1):187–196, 1997.

[13] Oana-Teodora Chis, Alejandro F. Villaverde, Julio R. Banga, and Eva Balsa-Canto. On the relationship between sloppiness and identifiability. *Mathematical Biosciences*, 282(Complete):147–161, 2016.

[14] Gilles Clermont and Sven Zenker. The inverse problem in mathematical biology. *Mathematical Biosciences*, 260:11–15, 2015.

[15] Claudio Cobelli and Joseph J. DiStefano III. Parameter and structural identifiability concepts and ambiguities: a critical review and analysis. *American Journal of Physiology*, 239(1):R7–R24, 1980.

[16] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley, 2012.

[17] Bryan C Daniels, Yan-Jiun Chen, James P Sethna, Ryan N Gutenkunst, and Christopher R Myers. Sloppiness, robustness, and evolvability in systems biology. *Current Opinion in Biotechnology*, 19(4):389 – 395, 2008.

[18] John Duchi. Derivations for linear algebra and optimization. 2007.

[19] Kamil Erguler and Michael P H Stumpf. Practical limits for reverse engineering of dynamical systems: a statistical analysis of sensitivity and parameter inferability in systems biology models. *Molecular BioSystems*, 7(5):1593–1602, May 2011.

[20] R. A. Fisher. Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 22(5):700?725, 1925.

[21] Benjamin L Francis and Mark K Transtrum. Unwinding the model manifold: choosing similarity measures to remove local minima in sloppy dynamical systems. *arXiv preprint arXiv:1805.12052*, 2018.

[22] Matan Gavish and David L. Donoho. The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Trans. Inform. Theory*, 60(8):5040–5053, 2014.

[23] Elizabeth Gross, Brent Davis, Kenneth L. Ho, Daniel J. Bates, and Heather A. Harrington. Numerical algebraic geometry for model selection and its application to the life sciences. *J. R. Soc. Interface*, 13, 2016.

[24] Ryan N. Gutenkunst, Jordan C. Atlas, Fergal P. Casey, Brian C. Daniels, Robert S. Kuczenski, Joshua J. Waterfall, Chris R. Myers, and James P. Sethna. Sloppycell, 2007.

[25] Ryan N. Gutenkunst, Joshua J. Waterfall, Fergal P. Casey, Kevin S. Brown, Christopher R. Myers, and James P. Sethna. Universally Sloppy Parameter Sensitivities in Systems Biology Models. *PLoS Computational Biology*, 3(10):e189–1878, oct 2007.

[26] Keegan E. Hines, Thomas R. Middendorf, and Richard W. Aldrich. Determination of parameter identifiability in nonlinear biophysical models: A Bayesian approach. *The Journal of general physiology*, 143(3):401–16, 2014.

[27] Serkan Hoşten, Amit Khetan, and Bernd Sturmfels. Solving the likelihood equations. *Foundations of Computational Mathematics*, 5(4):389–407, 2005.

[28] Joseph DiStefano III. *Dynamic Systems Biology Modeling and Simulation, 1st Edition*. Elsevier: Academic Press, Amsterdam, 2013.

[29] M. Joshi, A. Seidel-Morgenstern, and A. Kremling. Exploiting the bootstrap method for quantifying parameter confidence intervals in dynamical systems. *Metabolic Engineering*, 8(5):447 – 455, 2006.

[30] Johan Karlsson, Milena Anguelova, and Mats Jirstrand. An efficient method for structural identifiability analysis of large dynamic systems*. *IFAC Proceedings Volumes*, 45(16):941 – 946, 2012.

[31] Gregor Kemper. Separating invariants. *Journal of Symbolic Computation*, 44(9):1212–1222, 2009.

[32] S. Kullback and R.A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 1951.

[33] Colin H LaMont and Paul A Wiggins. A correspondence between thermodynamics and inference. *arXiv preprint arXiv:1706.01428*, 2017.

[34] Daniel Lazard. Injectivity of real rational mappings: the case of a mixture of two Gaussian laws. *Math. Comput. Simulation*, 67(1-2):67–84, 2004.

[35] Lennart Ljung. *System identification: theory for the user*. Prentice Hall Information and System Sciences Series. Prentice Hall, Inc., Englewood Cliffs, NJ, 1987.

[36] Lennart Ljung and Torkel Glad. On global identifiability for arbitrary model parametrizations. *Automatica*, 30(2):265 – 276, 1994.

[37] Brian K. Mannakee, Aaron P. Ragsdale, Mark K. Transtrum, and Ryan N. Gutenkunst. Sloppiness and the geometry of parameter space. In Liesbet Geris and David Gomez-Cabrero, editors, *Uncertainty in Biology: A Computational Modeling Approach*, pages 271–299. Springer International Publishing, Cham, 2016.

[38] Gabriella Margaria, Eva Riccomagno, Michael J. Chappell, and Henry 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):1–26, 2001.

[39] Nicolette Meshkat, Chris Anderson, and Joseph J Distefano. Finding identifiable parameter combinations in nonlinear ODE models and the rational reparameterization of their input-output equations. *Mathematical Biosciences*, 233(1):19–31, sep 2011.

[40] Nicolette Meshkat, Marisa Eisenberg, and Joseph J. DiStefano, III. An algorithm for finding globally identifiable parameter combinations of nonlinear ODE models using Gröbner bases. *Mathematical Biosciences*, 222(2):61–72, 2009.

[41] Eva Balsa-Canto Oana-Teodora Chis, Julio R. Banga. Structural identifiability of systems biology models: A critical comparison of methods. *PLoS ONE*, 6(11), 2011.

[42] François. Ollivier. *Le Problème de l'Identifiabilité Structurelle Globale: Étude Théorique, Méthodes Effectives et Bornes de Complexité.* PhD thesis, École Polytéchnique, 1990.

[43] Dhruva V. Raman, James Anderson, and Antonis Papachristodoulou. On the performance of nonlinear dynamical systems under parameter perturbation. *Automatica*, 63:265 – 273, 2016.

[44] Dhruva V. Raman, James Anderson, and Antonis Papachristodoulou. Delineating parameter unidentifiabilities in complex models. *Phys. Rev. E*, 95:032314, Mar 2017.

[45] C Radhakrishna Rao. Information and accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc*, 37(3):81–91, 1945.

[46] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–9, 2009.

[47] Andreas Raue, Johan Karlsson, Maria Pia Saccomani, Mats Jirstrand, and Jens Timmer. Comparison of approaches for parameter identifiability analysis of biological systems. *Bioinformatics (Oxford, England)*, 30(10):1440–8, 2014.

[48] Thomas J. Rothenberg. Identification in parametric models. *Econometrica*, 39(3):577–591, 1971.

[49] Alexandre Sedoglavic. A probabilistic algorithm to test local algebraic observability in polynomial time. In *Proceedings of the International Symposium on Symbolic and Algebraic Computation*. ACM, 2001.

[50] Jim Sethna. Fitting Polynomials: Where is sloppiness from? Webpage last modified June 11, 2008, `http://www.lassp.cornell.edu/sethna/Sloppy/FittingPolynomials.html`.

[51] Daniel Silk, Paul D W Kirk, Christopher P Barnes, Tina Toni, and Michael P H Stumpf. Model selection in systems biology depends on experimental design. *PLOS Computational Biology*, 10(6), 2014.

[52] E. D. Sontag. For differential equations with r parameters, $2r + 1$ experiments are enough for identification. *Journal of Nonlinear Science*, 12(6):553–583, 2002.

[53] Seth Sullivant. Algebraic Statistics. In preparation, `http://www4.ncsu.edu/~smsulli2/Pubs/asbook.html`.

[54] Christian Tönsing, Jens Timmer, and Clemens Kreutz. Cause and cure of sloppiness in ordinary differential equation models. *Phys. Rev. E*, 90:023303, Aug 2014.

[55] Mark K. Transtrum, Benjamin B. Machta, Kevin S. Brown, Bryan C. Daniels, Christopher R. Myers, and James P. Sethna. Perspective: Sloppiness and emergent theories in physics, biology, and beyond. *The Journal of Chemical Physics*, 143(1):010901, jul 2015.

[56] Mark K. Transtrum, Benjamin B. Machta, and James P. Sethna. Why are nonlinear fits to data so challenging? *Physical Review Letters*, 104(6):060201, feb 2010.

[57] Mark K. Transtrum, Benjamin B. Machta, and James P. Sethna. Geometry of nonlinear least squares with applications to sloppy models and optimization. *Physical Review E*, 83(3):036701, mar 2011.

[58] Mark K. Transtrum and Peng Qiu. Model reduction by manifold boundaries. *Physical Review Letters*, 113(9):098701, aug 2014.

[59] Sandor Vajda, Herschel Rabitz, Eric Walter, and Yves Lecourtier. Qualitative and quantitative identifiability analysis of nonlinear chemical kinetic models. *Chemical Engineering Communications*, 83(1):191–219, 1989.

[60] Michele Vallisneri. Use and abuse of the fisher information matrix in the assessment of gravitational-wave parameter-estimation prospects. *Phys. Rev. D*, 77:042001, Feb 2008.

[61] Joshua J. Waterfall, Fergal P. Casey, Ryan N. Gutenkunst, Kevin S. Brown, Christopher R. Myers, Piet W. Brouwer, Veit Elser, and James P. Sethna. The sloppy model universality class and the Vandermonde matrix. *Physical Review Letters*, 97(15):150601, 2006.

A. Table summary of main examples

Example	Fitting points to a line 2.5 2.133.104.8
Type	Explicit time-dependent model
Parameter space P	$(a_0, a_1) \in \mathbb{R}^2$
Variable ($x \in X$)	$x(t) \in \mathbb{R}$ for $t \in \mathbb{R}_{\geq 0}$
Measurable output ($y \in Y$)	$x(t)$
Perfect data	$(x(t_1), \ldots, x(t_N))$ for some $t_1 < \cdots < t_N \in \mathbb{R}_{\geq 0}$
Noisy data	$(x(t_1) + \epsilon_1, \ldots, x(t_N) + \epsilon_N)$

Example	Two biased coins 2.1, 2.1
Type	Finite discrete statistical model
Parameter space P	$[0,1]^3$
Variable ($x \in X$)	Outcome of 1 instance of the experiment
Measurable output ($y \in Y$)	Record of 1 instance of the experiment
Perfect data	Probability distribution for $p \in P$
Noisy data	Record of N instances of the experiment

Example	Sum of exponentials 2.6, 2.11, 3.11, 4.14
Type	Explicit time dependant model
Parameter space P	$(a,b) \in \mathbb{R}^2_{\geq 0}$
Variable ($x \in X$)	$x(t) \in \mathbb{R}_{>0}$ for $t \in \mathbb{R}_{\geq 0}$
Measurable output ($y \in Y$)	$x(t) \in \mathbb{R}_{>0}$ for $t \in \mathbb{R}_{\geq 0}$
Perfect data	$(x(t_1), \ldots, x(t_N))$ for some $t_1 < \cdots < t_N \in \mathbb{R}_{\geq 0}$
Noisy data	$(x(t_1) + \epsilon_1, \ldots, x(t_N) + \epsilon_N)$

Example	An ODE model with an exact solution 2.7
Type	Polynomial ODE model
Parameter space P	$(p_1, p_2) \in \mathbb{R}^2_{>0}$
Variable ($x \in X$)	$(x_1(t), x_2(t)) \in \mathbb{R}^2_{>0}$ for $t \in \mathbb{R}_{\geq 0}$
Measurable output ($y \in Y$)	$(x_1(t), x_2(t))$
Perfect data	$(x_1(t_1), x_2(t_1), \ldots, _1 (t_N), x_2(t_N))$ for some $t_1 < \cdots < t_N \in \mathbb{R}_{\geq 0}$
Noisy data	$(x_1(t_1) + \epsilon_{1,1}, x_2(t_1) + \epsilon_{2,1}, \ldots, _1 (t_N) + \epsilon_{1,N}, x_2(t_N) + \epsilon_{2,N})$

Example	A non-linear ODE model 2.15
Type	Polynomial ODE model
Parameter space P	$(p_1, p_2, p_3, p_4, p_5) \in P \subseteq \mathbb{R}^5$
Variable ($x \in X$)	$(x_1(t), x_2(t)) \in \mathbb{R}^2_{>0}$ for $t \in \mathbb{R}_{\geq 0}$
Measurable output ($y \in Y$)	$(x_1(t), x_2(t)) \in \mathbb{R}^2_{>0}$ for $t \in \mathbb{R}_{\geq 0}$
Perfect data	• $(x_1(t_1), x_2(t_1), \ldots, _1(t_N), x_2(t_N))$ for some $t_1 < \cdots < t_N \in \mathbb{R}_{\geq 0}$ • An exhaustive summary or input-output equations
Noisy data	$(x_1(t_1) + \epsilon_{1,1}, x_2(t_1) + \epsilon_{2,1}, \ldots, _1(t_N) + \epsilon_{1,N}, x_2(t_N) + \epsilon_{2,N})$

Example	Gaussian mixtures 3.9
Type	Continuous parametric statistical model
Parameter space P	$(\lambda, \mu, \sigma, \nu, \tau) \in [0,1] \times \mathbb{R} \times \mathbb{R}_{\geq 0} \times \mathbb{R} \times \mathbb{R}_{\geq 0}$
Variable ($x \in X$)	a characteristic $x \in \mathbb{R}_{\geq 0}$ of a mixed polulation
Measurable output ($y \in Y$)	$x \in \mathbb{R}_{\geq 0}$
Perfect data	• Probability distribution of x for some $(\lambda, \mu, \sigma, \nu, \tau)$ • Value of the cdf (or the pdf) at general $x_1, \ldots, x_1 1 \in \mathbb{R}$ • Value of the moment generating function at general $t_1, \ldots, t_1 1 \in (-a, a)$ • 11 generic moments (or the first 7)
Noisy data	• Measurements from a finite sample with some measurement error • Empirical distribution function from a finite sample (or its numerical derivative) • 11 generic sample moments from a finite sample (or the first 7)

Example	Linear parameter-varying model 4.10
Type	ODE model
Parameter space P	$p \in \mathbb{R}^{r-1}$
Variable ($x \in X$)	$x(t) \in \mathbb{R}^m$ for $t \in \mathbb{R}_{\geq 0}$
Measurable output ($y \in Y$)	$y = Cx(t)$ for $t \in \mathbb{R}_{\geq 0}$
Perfect data	$(y(t_1), \ldots, y(t_N))$ for some $t_1 < \cdots < t_N \in \mathbb{R}_{\geq 0}$
Noisy data	$(y(t_1) + \epsilon_1, \ldots, y(t_N) + \epsilon_N)$