Identifiability and Reconstruction of Biochemical Reaction Networks from Population Snapshot Data

Authors:

Eugenio Cinquemani

Date Submitted: 2019-04-08

Keywords: flow-cytometry, reporter gene systems, statistical inference, regulatory networks, moment equations

Abstract:

Inference of biochemical network models from experimental data is a crucial problem in systems and synthetic biology that includes parameter calibration but also identification of unknown interactions. Stochastic modelling from single-cell data is known to improve identifiability of reaction network parameters for specific systems. However, general results are lacking, and the advantage over deterministic, population-average approaches has not been explored for network reconstruction. In this work, we study identifiability and propose new reconstruction methods for biochemical interaction networks. Focusing on population-snapshot data and networks with reaction rates affine in the state, for parameter estimation, we derive general methods to test structural identifiability and demonstrate them in connection with practical identifiability for a reporter gene in silico case study. In the same framework, we next develop a two-step approach to the reconstruction of unknown networks of interactions. We apply it to compare the achievable network reconstruction performance in a deterministic and a stochastic setting, showing the advantage of the latter, and demonstrate it on population-snapshot data from a simulated example.

Record Type: Published Article

Submitted To: LAPSE (Living Archive for Process Systems Engineering)

Citation (overall record, always the latest version):	LAPSE:2019.0446
Citation (this specific file, latest version):	LAPSE:2019.0446-1
Citation (this specific file, this version):	LAPSE:2019.0446-1v1

DOI of Published Version: https://doi.org/10.3390/pr6090136

processes

MDPI

Article

Identifiability and Reconstruction of Biochemical Reaction Networks from Population Snapshot Data

Eugenio Cinquemani

Univ. Grenoble Alpes, Inria, 38000 Grenoble, France; eugenio.cinquemani@inria.fr

Received: 28 June 2018; Accepted: 15 August 2018; Published: 22 August 2018

Abstract: Inference of biochemical network models from experimental data is a crucial problem in systems and synthetic biology that includes parameter calibration but also identification of unknown interactions. Stochastic modelling from single-cell data is known to improve identifiability of reaction network parameters for specific systems. However, general results are lacking, and the advantage over deterministic, population-average approaches has not been explored for network reconstruction. In this work, we study identifiability and propose new reconstruction methods for biochemical interaction networks. Focusing on population-snapshot data and networks with reaction rates affine in the state, for parameter estimation, we derive general methods to test structural identifiability and demonstrate them in connection with practical identifiability for a reporter gene in silico case study. In the same framework, we next develop a two-step approach to the reconstruction of unknown networks of interactions. We apply it to compare the achievable network reconstruction performance in a deterministic and a stochastic setting, showing the advantage of the latter, and demonstrate it on population-snapshot data from a simulated example.

Keywords: moment equations; regulatory networks; statistical inference; reporter gene systems; flow-cytometry

1. Introduction

A central problem in systems and synthetic biology is the calibration of the unknown parameters of a biochemical reaction network model on the basis of real data [1]. Sometimes, the network of interactions in the system of interest is not completely known and needs to be reconstructed from data as well [2]. In synthetic biology, this is of special interest since the engineering of circuits into cells may result in unexpected interactions [3]. For parameter inference, a central question is what parameters of a given model can be identified unambiguously by a given experimental setup. This is the problem of identifiability, which is dedicated much attention in the systems' biology literature since crucial for the correct interpretation of inference results [4–6]. Posed purely in terms of model properties, the question goes under the name of structural identifiability. When instead the quality of the data is also taken into account, attention shifts to estimation accuracy and leads to the concept of practical identifiability. With an appropriate reformulation, similar questions can be addressed for the reconstruction of interaction networks.

Model experimental technologies in cellular biology allow one to measure reaction dynamics with single-cell resolution [7]. Correspondingly, stochastic reaction network models are utilized to explain variability of reaction dynamics, notably gene expression, across cells. It has been realized that the combination of stochastic modelling and single-cell measurements allows one to discriminate parameters of reaction models that would not be identifiable based on deterministic modelling counterparts and population-averaged data [8–10]. While demonstrated on specific case studies, this point has not been addressed in much generality. At the same time, one expects that the advantage of stochastic modelling and single-cell measurements for parameter identifiability carries over to the

problem of network inference. Whereas network inference has received attention in a deterministic context [11,12], the problem has barely been addressed in a stochastic setting, and a comparative study of network identifiability in the two settings does not exist.

In this paper, we discuss identifiability and reconstruction of unknown parameters and interactions in biochemical reaction networks. We focus on the fundamental class of models with reaction rates that are affine functions of the network state, and look at inference problems from mean and covariance time-profiles of the network state, as obtained from population-snapshot measurements [13]. State-affine rates reflect networks composed of reactions of zeroth or first order. Such networks occur naturally in gene expression modelling [14] and constitute the starting point to investigate more complex networks [15]. Population snapshot data, on the other hand, is easily obtained e.g., by inexpensive cytometry measurements, but is also the most immediate outcome of videomicroscopy [7]. Thus, we work in a framework that is at the same time sufficiently flexible and amenable of in-depth analysis via, in particular, the moment equations [16,17]. This allows us to establish the following tools and results.

For the parameter estimation problem, we develop general results and inexpensive methods to test structural identifiability, and discuss connections with practical identifiability. In the case of a random telegraph model of gene expression [14], we apply our methods to extend existing identifiability results [8,10] and discuss practical identifiability on the basis of numerical simulations. For the reconstruction of interaction networks, we develop a two-step method where a first step devoted to the identification of the network moment dynamics is followed by a second step that determines the network of interactions by the algorithmic solution of a factorization problem. We notably compare the deterministic (population-average) and stochastic (single-cell) scenarios, demonstrating the superiority of the latter from a theoretical viewpoint and also numerically on a toy example.

This work is an extension and consolidation of earlier results by the same author and published as part of conference proceedings in [18] (parameter identifiability) and [19] (network reconstruction). Relevant work on parameter identifiability in the context of biochemical networks has been developed by several authors (see [4,6,8,20] and references therein). Relative to the state of the art, our results on structural identifiability are entirely novel, while the discussion of practical identifiability is in the same spirit as e.g., [1]. Reconstruction of biochemical networks is reviewed in [11,12,21,22] (see also [23]) and is even dedicated a yearly challenge boosting continuous investigation (see [2]). However, from the viewpoint of stochastic dynamics and single-cell data, existing contributions are essentially limited to model discrimination and selection among small pools of specific network alternatives (see, e.g., [24]). Moreover, to our knowledge, a comparative analysis of network identifiability in the deterministic and stochastic setups is not present in literature. Our work contributes to fill this gap with an original analysis and methods for the reconstruction of stochastic biochemical interaction networks with no a priori restriction on the network structure for a given set of species.

The paper is organized as follows. In Section 2, we review stochastic modelling of biochemical reaction networks. In particular, in Section 2.1, we derive an especially convenient form for the moment equations that will be the working tool for the rest of the work. Section 3 is devoted to parameter identifiability. In particular, our novel methods for structural identifiability are derived in Section 3.1. Together with the practical identification tools reviewed in Section 3.2, they are applied to the case of the random telegraph model for a reporter gene system in Section 3.3. Section 4 addresses network reconstruction. The theoretical analysis and development of the two steps (Sections 4.1 and 4.2) of the reconstruction problem are turned into a novel practical network identification procedure in Section 4.3. Analysis and reconstruction methods are then demonstrated on a toy network in Section 4.4. Final discussion is presented in Section 5. For mathematical formulas, we stick to common notation, adding explanations where needed.

2. Modelling of Biochemical Reaction Networks

A reaction network is a family of $n \in \mathbb{N}$ chemical species $\mathscr{S}_1, \ldots, \mathscr{S}_n$ and $m \in \mathbb{N}$ reactions $\mathscr{R}_1, \ldots, \mathscr{R}_m$ that may occur among them in a given reaction volume. The stoichiometry matrix of the network, $S = [S_1, \ldots, S_m] \in \mathbb{Z}^{n \times m}$, is defined such that the ith row of S_j is the net change in the number of molecules of $\mathscr{S}_i$ when reaction $\mathscr{R}_j$ occurs, with $i = 1, \ldots, n$ and $j = 1, \ldots, m$. Let $X_i(t)$ be the number of molecules of $\mathscr{S}_i$ at time t, and $X(t) = [X_1(t) \ \cdots \ X_n(t)]^T$. The evolution of $X(t)$ depends on the random occurrence of reactions $\mathscr{R}_j$, thus $X(t)$ is, in general, a random vector. Under suitable assumptions on the reaction volume and the physics of the reactions [25], one has that $X(t)$ follows the laws of a Continuous-Time Markov Chain (CTMC), with the columns of S as possible state transitions and transition probabilities

$$\mathbb{P}[X(t+dt) = x + S_j | X(t) = x] = a_j(x,t)dt + o(dt), \qquad j = 1, \ldots, m.$$

Reaction rates $a(x,t) = [a_1(x,t), \ldots, a_m(x,t)]^T$ depend on the current abundance x of molecules of the different species, and possibly on time t e.g., in the presence of environmental perturbations. They entirely specify the time dynamics of $\mathbb{P}[X(t)]$ as expressed by the Chemical Master Equation [25].

We focus on reaction rates that are affine in the state, that is,

$$a(x,t) = Wx + w_0(t), \tag{1}$$

where $W \in \mathbb{R}^{m \times n}$ and $w_0(t) : \mathbb{R} \to \mathbb{R}^m$ is piecewise continuous (W may as well depend on t, but we will not address this scenario here). This form is typical of networks comprising zeroth- and first-order reactions only, as dictated by the mass-action laws [26]. In cellular biology, affine rates arise naturally in the modelling of gene expression (see later Section 3.3), and constitute the starting point for the (possibly approximate) modelling and analysis of regulatory networks [15]. The choice of state-affine rates and possible generalizations are further discussed in Section 5.

Note that, together, the patterns of nonzero entries of S and W define the network of regulatory interactions among the different species (see also [27]). Indeed, for every reaction j, the jth row of W tells what species regulate the reaction rate, whereas the jth column of S tells what species are affected by that reaction. Thus, if $S_{i,j} \neq 0$ and $W_{j,k} \neq 0$ for some j, X_k directly regulates the dynamics of X_i, whereas, if $S_{i,j} = 0$ or $W_{j,k} = 0$ for all j, X_k does not exert a direct regulation of the dynamics of X_i.

Denote $\mu(t) = \mathbb{E}[X(t)]$ and $\Sigma(t) = \mathrm{Var}(X(t)) = \mathbb{E}[(X(t) - \mu(t))(X(t) - \mu(t))^T]$ ("$\mathbb{E}[\cdot]$" indicates expectation). It can be shown that the time evolution of the mean vector μ and of the covariance matrix Σ obeys the so-called Moment Equations [16],

$$\dot{\mu}(t) = SW\mu(t) + Sw_0(t), \tag{2}$$

$$\dot{\Sigma}(t) = SW\Sigma(t) + \Sigma(t)W^T S^T + S\mathrm{diag}\big(W\mu(t) + w_0(t)\big)S^T, \tag{3}$$

with initial conditions $\mu(t_0) = \mu_0$ and $\Sigma(t_0) = \Sigma_0$ determined by the initial probability distribution of X at a time t_0 (we will conventionally fix this time to $t_0 = 0$). This is a set of linear equations in the entries of μ and Σ. Similar, though more intricate, equations could be written for higher-order moments [17], but we will not use them. Equation (2), describing the evolution of the process mean, also provides by definition the mean-field approximation of the system and, for networks with negligible random fluctuations, its deterministic dynamics. Equation (3) instead quantifies the strength of the stochastic fluctuations around the system mean. If rates are not in the form (1), the description of the moments and the mean-field approximation of the system dynamics are involved (see, e.g., [28,29]).

2.1. Vector Representation of Moment Equations

Thanks to linearity, Equations (2) and (3) can be rewritten in the form

$$\dot{z}(t) = Az(t) + Bw_0(t), \qquad t \geq 0, \tag{4}$$

where $z(t)$ is a vector formed by the entries of $\mu(t)$ and $\Sigma(t)$, whereas A and B are matrices of appropriate dimension fixed by S and W. This representation is non-unique as it depends in particular on the ordering of the entries of z. One such representation is the following. Let "vec$(\cdot)$" be the operation of stacking the columns of a matrix into a single column vector, and let "$\otimes$" denote the Kronecker product.

Proposition 1. *Let* $\underline{\Sigma} \triangleq \text{vec}(\Sigma) \in \mathbb{R}^{n^2}$ *and* $S^{(2)} \triangleq \begin{bmatrix}\text{vec}(S_1 S_1^T) & \cdots & \text{vec}(S_m S_m^T)\end{bmatrix}$. *Equation* (4) *holds with*

$$z = \begin{bmatrix}\mu(t) \\ \underline{\Sigma}(t)\end{bmatrix}, \qquad A = \begin{bmatrix}SW & 0_{n\times n^2} \\ S^{(2)}W & I_n \otimes (SW) + (SW) \otimes I_n\end{bmatrix}, \qquad B = \begin{bmatrix}S \\ S^{(2)}\end{bmatrix}. \tag{5}$$

Proof. We start by rewriting (3) in vector form. Let us drop index t from notation for brevity. By the properties of Kronecker product, one gets that

$$\dot{\underline{\Sigma}} = [I_n \otimes (SW) + (SW) \otimes I_n] \cdot \underline{\Sigma} + \text{vec}(S\text{diag}(W\mu + w_0)S^T).$$

In order to write the rightmost term in a more convenient form, observe that

$$S\text{diag}(W\mu + w_0)S^T = \sum_{j=1}^{n} S\text{diag}(W_j)S^T\mu_j + S\text{diag}(w_0)S^T,$$

where W_j denotes the j-th column of W. Therefore,

$$\text{vec}(S\text{diag}(W\mu + w_0)S^T) = \left[\text{vec}(S\text{diag}(W_1)S^T)\ ,\ \ldots\ ,\ \text{vec}(S\text{diag}(W_n)S^T)\right] \cdot \mu + \text{vec}(S\text{diag}(w_0)S^T).$$

Next note that, for any vector w of appropriate size, $\text{vec}(S\text{diag}(w)S^T) = S^{(2)}w$. Then, we may write

$$\dot{\underline{\Sigma}} = [I_n \otimes (SW) + (SW) \otimes I_n] \cdot \underline{\Sigma} + S^{(2)}W \cdot \mu + S^{(2)} \cdot w_0.$$

Together with Equation (2), this yields the result. □

In the sequel, we let (5) be the definition of z, A and B. Note that this representation is redundant, since the upper- and lower-triangular part of the symmetric matrix Σ are both included in z. In addition, depending on the definition of process $X(t)$, relationships among the entries μ and Σ may further reduce the effective dimensionality of the system [30].

2.2. Input–Output Model

For identification purposes, well-defined time-varying stimuli are usually applied to a biochemical network to explore its dynamics. These perturbation inputs typically enter the system in a way that is known only in part. Here, we focus on the scenario where the inputs act on rates $w_0(t)$. A natural manner to express the dependency of $w_0(t)$ on known inputs $u(t)$ is to write

$$w_0(t) = Gu(t), \tag{6}$$

where G is an $m \times n_u$ matrix of (possibly null) constants, while $u(t)$ is an n_u-dimensional vector of (possibly constant) functions. In general, we let $u(\cdot)$ belong to the space of piecewise-continuous

(vector) functions $\mathscr{U}$. In the sequel, we will consider $u(\cdot)$ to be known and treat G as possibly unknown parameters, reflecting partially unknown effects of inputs of reaction rates.

A chief biological application of stochastic reaction networks is to describe variability of the individual cell dynamics in an isogenic population. In this context, the different random outcomes of $X(t)$ correspond to different dynamics of individual cells in the same environment. Accordingly, Equations (2) and (3) describe how the statistics of $X(t)$ over the cell population evolve over time. Depending on the biochemical system under consideration, several experimental techniques are available to monitor the evolution of these statistics, ranging from population-average to individual-cell measurements. From the mathematical viewpoint, they differ in the order of the moments and in the state variables that are monitored. As far as moments up to second order are concerned, we may describe the observed moments as

$$y(\cdot) = C \cdot z(\cdot), \tag{7}$$

where C is a matrix of size $n_y \times (n + n^2)$ and z obeys Equation (4). Typically C is a $(0, 1)$-matrix that selects the observed entries of z, so that n_y is simply the number of observed moments. For a sequence of N_y measurement times t_ℓ, with $\ell = 1, \ldots, N_y$, we may then define the experimental measurements of y as

$$\tilde{y}_\ell = y(t_\ell) + e_\ell, \qquad \ell = 1, \ldots, N_y, \tag{8}$$

where e_ℓ accounts for measurement error.

For population-average measurements, C only selects the entries of y from those of μ. Note that the selected entries of μ, corresponding to different species $\mathscr{S}_i$, can be observed in separate experiments, as long as the moment equations (that is, the network state statistics) are the same across all experiments. For individual-cell measurements obtained e.g., by flow-cytometry, C most often selects the entries of μ and $\mathrm{diag}(\Sigma)$ in accordance with the monitored species $\mathscr{S}_i$. Again, several species monitored in different experiments are accounted for at once by Equation (8) as long as z obeys the same moment equations across all experiments. A definition of C such that non-diagonal entries $\Sigma_{i,i'}$ of Σ enter vector y instead subsumes experiments where species $\mathscr{S}_i$ and $\mathscr{S}_{i'}$ are quantified simultaneously in individual cells (covariance of X_i and $X_{i'}$ cannot be determined from independent experiments separately targeting i and i'). In general, for the later mathematical developments, C can be any real matrix of appropriate size. Finally, the statistical model for the error terms e_ℓ depends on the experimental technique. We will not discuss different possible error models here. Where needed in the sequel, following a common practice, we will assume the e_ℓ to be mutually independent Gaussian random vectors,

$$e_\ell \sim \mathscr{N}(0, R_\ell^2), \qquad \mathbb{E}[e_\ell e_{\ell'}] = 0, \ \ell \neq \ell', \tag{9}$$

with the R_ℓ^2 known (or estimated from the data). In connection with the discussion above, this model is notably advocated in [9] for flow-cytometry measurements.

In view of the above definitions, in vector form, the moment equation system relating inputs u to the observed outputs y is

$$\dot{z}(t) = Az(t) + BGu(t), \tag{10}$$

$$y(t) = Cz(t), \tag{11}$$

with z, A and B as in Equation (5), and time-sampled, noisy measurements of y as in Equations (8) and (9).

3. Identification of Parameters

Assuming S and C known, in this section, we consider the problem of estimating the reaction rate parameters W and G that define the model of Equations (10) and (11) via (5) from experimental

data (8) and (9). Since μ_0 and Σ_0 are also unknown in most practical scenarios, we let $z_0 = z(0)$ be an additional parameter vector to be estimated. We assume that $u(t)$ is a known input vector of the reaction network, such as controlled environmental stimuli. The choice of appropriate inputs and observables, i.e., function $u(\cdot)$ and matrix C, in an experiment design phase is an intriguing and important subject, but its full treatment is beyond the scope of this paper. We will limit ourselves to commenting on this point when relevant.

In general, some entries of W, G and z_0, such as null entries for species that do not participate in certain reactions or that are initially absent, may be known in advance. At the same time, as in the example of Section 3.3 later on, some of the unknown entries of W, G and z_0 may be identical by construction. Taking this into account, let $\theta = [\theta_1, \ldots, \theta_N]^T$ be the vector of unknown, distinguished entries of W, G, and z_0, in some convenient order. We assume that $\theta \in \Theta$, where $\Theta \subseteq \mathbb{R}^N$ is given. To avoid technical complicacies, we let Θ be nonempty and open, and elaborate on this hypothesis when appropriate.

The first question to be addressed is whether the values of θ can be uniquely resolved based on the system observables, regardless of the quality and frequency of the measurements. This is the problem of structural identifiability, which concerns properties of the system model and its observables. In our context, given model (10)–(11), it addresses the question whether the relationship between θ and $y(\cdot)$ is one-to-one. Structural identifiability is treated in Section 3.1.

The second question is the accuracy by which the model parameters can be estimated from actual data. This is the problem of practical identifiability. In addition to the model properties, it also concerns the quality and frequency of the data, as defined by Equation (8), and the properties of the measurement error. Intuitively, this question is only well-posed for systems that are structurally identifiable. However a formal definition of practical identifiability is not obvious and still open (see [1,4,6,20], among others). Rather than a theoretical discussion of practical identifiability, we will briefly review the Maximum Likelihood (ML) approach to identification from noisy data and related performance evaluation tools in Section 3.2, and discuss practical identifiability results in relation with estimation performance and structural identifiability on a numerical case study in Section 3.3.

3.1. Structural Identifiability

In this section, we review the formal notion of structural identifiability and develop novel methods to test structural identifiability for networks with affine rates. For any given $\theta \in \Theta$, let $\hat{y}_\theta$ be the corresponding solution $y(t)|_{t\geq 0}$ of Equations (10) and (11), and let

$$\mathscr{M}(\Theta) = \{\hat{y}_\theta : \theta \in \Theta\}. \tag{12}$$

Structural identifiability is then a property of the model family (12), as expressed via the following classical definitions [31].

Definition 1 (Identifiability at a point)**.** *The model family* (12) *is*

(a) locally identifiable at $\theta^* \in \Theta$ *if, for some neighborhood* $B_{\theta^*} \subseteq \Theta$ *of* θ^*,

$$\hat{y}_\theta = \hat{y}_{\theta^*} \implies \theta = \theta^*, \quad \forall \theta \in B_{\theta^*};$$

(b) globally identifiable at θ^* *if the implication above holds for* $B_{\theta^*} = \Theta$.

Definition 2 (Structural identifiability)**.** *The model family* (12) *is* locally identifiable *(resp.* globally identifiable*) if the property (a) (resp. (b)) of Definition 1 is true for almost every (a.e.)* $\theta^* \in \Theta$.

(Notice that, for this definition, the hypothesis that set Θ is open is irrelevant, since a property holding almost everywhere in an open set also holds almost everywhere in its closure.) Structural identifiability can thus be understood in a global or local sense. In both cases, an equivalent formulation

of structural identifiability can be given in terms of the Laplace transform of $\hat{y}_\theta$, which we denote by $Y(s,\theta)$, with $s \in \mathbb{C}$. Indeed, the model family in (12) is structurally identifiable whenever, for a.e. $\theta^* \in \Theta$,

$$Y(\cdot,\theta) = Y(\cdot,\theta^*) \implies \theta = \theta^* \tag{13}$$

holds for θ within some B_{θ^*} (local identifiability) or for all $\theta \in \Theta$ (global identifiability) [31]. Global identifiability is generally difficult to assess (we will discuss why by the case study of Section 3.3). Local identifiability can instead be tested on the basis of the following approach. Let

$$\nabla Y(s,\theta) \triangleq \frac{\partial Y}{\partial \theta}(s,\theta) = \left[\frac{\partial Y}{\partial \theta_1} \cdots \frac{\partial Y}{\partial \theta_N}\right](s,\theta).$$

Proposition 2. *Suppose that, for some $L \in \mathbb{N}$, a set of points $\mathscr{S}_L = \{s_1,\ldots,s_L\} \subseteq \mathbb{R}$ exists such that the matrix*

$$\Delta(\mathscr{S}_L,\theta^*) \triangleq \begin{bmatrix} \nabla Y(s_1,\theta^*) \\ \vdots \\ \nabla Y(s_L,\theta^*) \end{bmatrix} \tag{14}$$

has full column rank. Then (12) is locally identifiable at θ^*.

Proof. In the hypothesis of full column rank, $\Delta(\mathscr{S}_L,\theta^*)\cdot(\theta-\theta^*)$ is not null for any $\theta \neq \theta^*$. Hence, for at least some $s \in \mathscr{S}$ and $\theta - \theta^*$ sufficiently small, $Y(s,\theta) \simeq Y(s,\theta^*) + \nabla Y(s,\theta)\cdot(\theta-\theta^*) \neq Y(s,\theta^*)$. That is, for θ in some ball B_{θ^*} around θ^*, $\theta \neq \theta^*$ implies $Y(\cdot,\theta) \neq Y(\cdot,\theta^*)$, which is equivalent to (13). □

This result, which is a variant of a family of rank conditions [32], provides a test for local identifiability at a given θ^* in the sense of Definition 1(a), and applies to any $Y(\cdot,\theta)$ differentiable with respect to θ. For the moment equations of our interest, it is easy to see that

$$Y(s,\theta) = C\big(sI - A(W)\big)^{-1}\big(z_0 + BGU(s)\big), \tag{15}$$

where $U(s)$ is the Laplace transform of $u(t)|_{t\geq 0}$, and the writing $A(W)$ indicates dependency of A on W. It can be verified by inspection that, for any fixed s, Equation (15) is a ratio of multivariate polynomials in the entries of W, G and z_0, that is, the entries of θ. This allows us to strengthen Proposition 2 into a method to test local structural identifiability as per Definition 2.

Corollary 1. *Given a set of points $\mathscr{S}_L$, if $\Delta(\mathscr{S}_L,\theta^*)$ has full column rank for at least one value of θ^*, then (12) is structurally locally identifiable.*

Proof. As a function of θ, rank loss of $\Delta(\mathscr{S}_L,\theta)$ relative to the rank of $\Delta(\mathscr{S}_L,\theta^*)$ may only occur at the zeroes of the minors of $\Delta(\mathscr{S}_L,\theta)$ that are nonzero for $\theta = \theta^*$. In view of (15), these minors are multivariate polynomials in the entries of θ. N-variate polynomials are either identically null, or their zeroes form a variety of zero measure relative to the N-dimensional Lebesgue measure of Θ, that is, they are nonzero almost everywhere. The nonzero minors of $\Delta(\mathscr{S}_L,\theta^*)$ are obviously not identically null, hence they are nonzero almost everywhere in Θ and guarantee that the rank of $\Delta(\mathscr{S}_L,\theta)$ is no smaller than the rank of $\Delta(\mathscr{S}_L,\theta^*)$ almost everywhere. Since by hypothesis $\Delta(\mathscr{S}_L,\theta^*)$ is full rank, $\Delta(\mathscr{S}_L,\theta)$ is full rank almost everywhere in Θ. By Proposition 2, one concludes that (12) is locally identifiable almost everywhere, i.e., it is structurally locally identifiable. □

Thus, the method suggested by Corollary 1 amounts to evaluate the rank of matrix $\Delta(\mathscr{S}_L,\theta^*)$ at suitable (even randomly generated) candidate points $\mathscr{S}_L$ and θ^*. One choice of $\mathscr{S}_L$ and θ^* yielding full rank suffices to conclude for structural identifiability.

A time-domain version of Proposition 2 can also be derived, resulting in the evaluation of the rank of the sensitivity matrix of the system outputs at sufficiently many time points (see e.g., [1] and references therein, and Equation (17) later on). Unfortunately, computation of time-domain sensitivities is entirely numerical (see [33]). We found that this makes the subsequent rank computation unreliable. Instead, for the moment equations of our concern, the calculation of the Laplace-domain sensitivity matrix $\Delta(\mathscr{S}_L, \theta^*)$ can be performed explicitly or by standard symbolic calculation software, since the matrix is composed of rational functions. This allows one to confine numerical rank evaluation at a very last step, with great gain in numerical stability. Therefore, not only does Corollary 1 establish a theoretical result showing that, for the model class (12), structural identifiability at a point implies structural stability *a.e.*, but it also provides a numerically robust method to test identifiability. We will further discuss implementation aspects in Section 3.3, where the method is demonstrated on a case study.

Note that, for fixed $\mathscr{S}_L$ and θ^*, the rank of $\Delta(\mathscr{S}_L, \theta^*)$ increases by the addition of rows, that is, by the observation of more system statistics (in (15), C has more rows). In particular, this suggests that observation of stochastic variability of a system (e.g., of $\Sigma(\cdot)$) favors identifiability of networks that are not identifiable from the sole observation of deterministic population-average dynamics (e.g., of $\mu(\cdot)$), in line with previous findings on specific case studies [8]. Another instance of this fact is the case study of Section 3.3.

To conclude the section, suppose that input $u(\cdot)$ can be selected from within the class of functions $\mathscr{U}$. The above identifiability definitions and results can be rephrased in terms of the existence of an element $u \in \mathscr{U}$ that makes (12) identifiable. In particular, the test of identifiability of Corollary 1 requires finding an element $u(\cdot)$ in $\mathscr{U}$ such that the corresponding $\Delta(\mathscr{S}_L, \theta^*)$ is full rank (a suitable input is often easy to determine by a qualitative inspection of the problem). We will come back on this point in the discussion of Section 5.

3.2. Parameter Identification in Practice

For a parametric model that is structurally identifiable, the question turns to how accurately the unknown parameters can be estimated in practice from real (noisy, sampled) data. In this section, we consider existing approaches to assess practical identifiability, which will be used in Section 3.3 to elaborate on the results of Section 3.1 on a case study, and later on in Section 4.3 to devise a network reconstruction algorithm. As argued in [20], by the very name, practical identifiability should be a property independent of the specific estimation strategy. However, from this perspective, the question is hard to tackle except for special cases, and a posteriori strategies have been popularized trying to assess estimation uncertainty around the estimated paramater values [6]. In this section, we focus on a widespread approach to parameter estimation, likelihood maximization, and discuss an equally common linearization approach to define estimation accuracy around the obtained parameter estimates [1,34].

Consider data (8). Under the statistical error model (9), the ML estimator of θ is defined as a vector $\hat{\theta}$ such that

$$\hat{\theta} \in \arg\min_{\theta \in \Theta} E(\theta), \qquad E(\theta) \triangleq \frac{1}{2} \sum_{\ell=1}^{N_y} \big(\tilde{y}_\ell - \hat{y}_\theta(t_\ell)\big)^T R_\ell^{-1} \big(\tilde{y}_\ell - \hat{y}_\theta(t_\ell)\big) \tag{16}$$

(neglecting terms independent of θ, the weighted sum of squares $E(\theta)$ equals the negative log-likelihood of θ given the data [35]). Solution $\hat{\theta}$ is unique (almost surely) as long as the finite sample set $\{y(t_\ell) : \ \ell = 1, \dots, N_y\}$ is sufficiently rich. Formally, this can be expressed by the condition that the sensitivity matrix

$$\begin{bmatrix} S_\theta(t_1) \\ \vdots \\ S_\theta(t_{N_y}) \end{bmatrix} \triangleq \begin{bmatrix} \frac{\partial \hat{y}_\theta(t_1)}{\partial \theta} \\ \vdots \\ \frac{\partial \hat{y}_\theta(t_{N_y})}{\partial \theta} \end{bmatrix} \tag{17}$$

has full rank (almost everywhere in Θ). In particular, this requires $N_y \cdot h \geq N$, where h is the size of y. As discussed in the previous section after Proposition 2, this coincides with the condition for structural identifiability expressed in the time domain.

ML estimators enjoy well-known, strong theoretical properties that are mostly asymptotic in N_y [35]. For a finite dataset, it is of interest to quantify the (approximate) statistics of the estimation error $\hat{\theta} - \theta^*$, where θ^* denotes the true value of θ that generated the data. Using the approximation $\hat{y}_\theta(t) \simeq \hat{y}_{\theta^*}(t) + S_{\theta^*}(t) \cdot (\theta - \theta^*)$, where S_{θ^*} is matrix (17) computed at $\theta = \theta^*$, one has that $\mathbb{E}[\hat{\theta}] = \theta^*$, and the estimation error covariance matrix $V(\theta^*) = \mathrm{Var}(\hat{\theta} - \theta^*)$ is equal to $\mathbb{I}(\theta^*)^{-1}$, where

$$\mathbb{I}(\theta^*) = \mathbb{E}\left[\frac{\partial^2 E(\theta)}{\partial\theta\partial\theta^T}\right]_{\theta=\theta^*} = \sum_{\ell=1}^{N_y} S_{\theta^*}(t_\ell)^T R_\ell^{-1} S_{\theta^*}(t_\ell).$$

Matrix $\mathbb{I}(\theta^*)$ is known as the Fisher Information Matrix (FIM, [1,4,6,34]). Provided $\hat{\theta}$ is reasonably close to θ^*, evaluating V at $\hat{\theta}$ allows one to establish a confidence ellipsoid $\mathscr{E}_\alpha(\hat{\theta})$ that θ^* belongs to with $1 - \alpha$ probability, $\alpha \in (0, 1)$. For R_ℓ known, with $\ell = 1, \ldots, N_y$, this is given by the set

$$\mathscr{E}_\alpha(\hat{\theta}) = \{\theta : \ (\theta - \hat{\theta})^T V(\hat{\theta})^{-1} (\theta - \hat{\theta}) \leq \chi_\alpha\}, \tag{18}$$

where χ_α is the $(1 - \alpha)$-quantile of the χ^2 distribution with N degrees of freedom. In practice, computation of V requires computation of (17). This is easily done by the sensitivity equations [33], a system of ODEs that follows from the definition of (10) to be solved numerically alongside the latter.

To conclude, following up from the final comments to the previous section, note that the choice of the perturbation input u (but also the observation matrix C) may play an important role in the achievable estimation performance. In particular, u should make the system structurally identifiable, and should be chosen such that V is as small as possible in an appropriate sense. This is the subject of optimal experiment design [31], which we will not treat here.

3.3. Example: Reporter Gene Expression Dynamics

By the tools developed in the previous sections, in this section, we study identifiability of reporter gene systems. These are synthetic gene constructs that are commonly engineered into cells to monitor gene expression over time in terms of the synthesis of a fluorescent (or luminescent) protein [36]. The illustration of a reporter gene is given in Figure 1. When the gene is expressed, the transcription of the genetic sequence that codes for the fluorescent protein leads to synthesis of corresponding mRNA molecules. These molecules are then transcribed in a second step, leading to the synthesis of fluorescent proteins. When the gene is switched off, synthesis of mRNA molecules is no longer possible until the gene is switched on again. The existing mRNA and protein molecules are subject to degradation throughout. Experimental measurement of fluorescence intensity allows for the quantification of the abundance of fluorescent reporter molecules over time, and thus gives an indirect dynamical readout of the gene expression state. In some cases, an additional maturation step of the reporter protein needs to be taken into account before the molecule becomes fluorescent. While this can be modelled in our mathematical framework, we will not address it here.

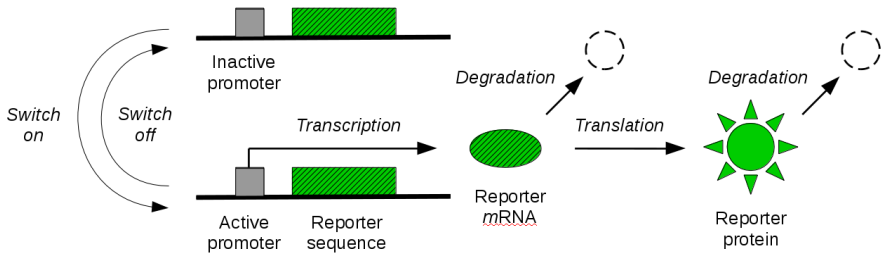


Figure 1. Reporter gene system. The coding sequence of a fluorescent reporter protein is engineered into a gene of interest. The gene promoter can switch between an inactive (off) and an active (on) state. When active, transcription of reporter *m*RNA molecules is enabled. Existing *m*RNA molecules are further translated into visible (quantifiable) reporter protein molecules. Both *m*RNA and protein molecules are subject to degradation.

Let M and P denote *m*RNA and protein species, in the same order. Let F be the active promoter species, such that gene expression is enabled only when F is present. The synthesis and degradation of *m*RNA and protein molecules described above are typically expressed by the reaction model

$$F \xrightarrow{k_M} F + M, \qquad M \xrightarrow{d_M} \varnothing, \tag{19}$$

$$M \xrightarrow{k_P} M + P, \qquad P \xrightarrow{d_P} \varnothing, \tag{20}$$

for some nonnegative rate parameters k_M, k_P, d_M and d_P [14,37]. In turn, for some nonnegative parameters λ_+ and λ_-, the switching dynamics of F are expressed by two additional reactions for activation and deactivation,

$$\varnothing \xrightarrow{r(\lambda_+)} F, \qquad F \xrightarrow{\lambda_-} \varnothing, \tag{21}$$

where the activation rate parameter $r(\lambda_+)$ is equal to λ_+ in the inactive state, and to 0 otherwise (activation is possible only from the inactive state).

At the level of single cells, Equations (19)–(21) are known as the random telegraph model of gene expression [14]. Let $X_1 \in \mathbb{N}$ and $X_2 \in \mathbb{N}$ be the number of molecules of species M and P, respectively. Let $X_3 \in \{0, 1\}$ be the state of the gene promoter, that is, $X_3 = 0$ in the inactive state and $X_3 = 1$ in the active state. In accordance with the modelling of Section 2, one may describe $X = [X_1\ X_2\ X_3]^T$ as a Markov process [14]. The rates of the $m = 6$ reactions (19)–(21) (ordered from left to right, top to bottom) are

$$a(x,t) = \begin{bmatrix} k_M x_3 & d_M x_1 & k_P x_1 & d_P x_2 & \lambda_+(1-x_3) & \lambda_- x_3 \end{bmatrix}^T.$$

These rates are in the affine form (1), with time-independent $w_0(t) = G$. Rate parameters and stoichiometry matrix of the network are

$$S = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}, \qquad W = \begin{bmatrix} 0 & 0 & k_M \\ d_M & 0 & 0 \\ k_P & 0 & 0 \\ 0 & d_P & 0 \\ 0 & 0 & -\lambda_+ \\ 0 & 0 & \lambda_- \end{bmatrix}, \qquad G = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \lambda_+ \\ 0 \end{bmatrix}.$$

We consider that all rate parameters of reactions (19)–(21), that is all parameters defining W and G, are unknown, and wish to study their identifiability from the experimentally observed species P. The observation model (7), that is, matrix C, is thus such that $y(t) = \left[\mathbb{E}[X_2(t)], \mathrm{var}(X_2(t))\right]^T$.

We first investigate structural identifiability by the approach of Section 3.1. This requires in the first place to compute the Laplace transform $Y(s,\theta)$ and its sensitivity function $\nabla Y(s,\theta)$. We implemented these computations in MATLAB (Release 2017b, The MathWorks, Inc., Natick, MA, USA) by the aid of the Symbolic calculation toolbox. Because $Y(s,\theta)$ is a rational (vector) function, this allows us to obtain explicit expressions for $\nabla Y(s,\theta)$ in the symbolic variables s and θ in a straightforward manner. Then, for a given set of test points $\mathscr{S}_L$ and a given vector θ^*, we evaluate $\nabla Y(s,\theta)$ numerically at $(s,\theta) = (s_\ell,\theta^*)$, for all values $s_\ell \in \mathscr{S}_L$. This allows us to build the sensitivity matrix $\Delta(\mathscr{S}_L,\theta^*)$ and finally compute its rank. In the light of Corollary 1, the system is found identifiable if this rank is full.

For our case study, given the number of test points L, matrix $\Delta(\mathscr{S}_L,\theta)$ has $2L$ rows. Therefore, the full column rank required by Corollary 1 to conclude for identifiability can only be obtained if $2L \geq N$, where N, the number of columns of $\Delta(\mathscr{S}_L,\theta)$, is equal to the size of the parameter vector θ. This is made of six rate parameters plus the unknown initial conditions z_0, for a total of $N = 15$ entries. Then, identifiability should be tested with $L \geq 8$. We arbitrarily set $L = 10$ and choose points $\mathscr{S}_L = \{s_l = l : \ l = 1,\ldots,L\}$. Then, we sampled values θ^* at random and computed the corresponding rank of $\Delta(\mathscr{S}_L,\theta^*)$. We always found this rank to be full, that is, we concluded in each case that the reporter gene system is structurally locally identifiable. Whereas one such random evaluation suffices, the consistency of the result shows the effectiveness of the method, which does not depend on a critical choice of θ^* or $\mathscr{S}_L$.

To verify how variance observations contribute to identifiability, by the same approach, we further investigated structural local identifiability from the observation of the reporter abundance (fluorescence) mean only. The method holds unchanged, provided a suitable redefinition of matrix C. In this case, $\Delta(\mathscr{S}_L,\theta)$ has only L rows, and its rank should be tested for $L \geq N$. For increasing values of L and parameter values θ^* sampled at random as above, the maximal rank of this reduced version of $\Delta(\mathscr{S}_L,\theta^*)$ was found to be equal to 7. That is, full column rank was never found. Since the rank condition of Corollary 1 is not necessary but only sufficient, we cannot state from this test that the reporter system lacks structural identifiability from the mean, yet the results are a strong hint toward non-identifiability. In fact, consider for simplicity $z_0 = 0$. The entry of $Y(s,\theta)$ associated with the mean is found to be

$$\lambda_+ k_M k_P / [s(\lambda_+ + \lambda_- + s)(d_M + s)(d_P + s)].$$

Inspection of this formula reveals lack of identifiability, since parameters k_M and k_P only appear through the product $k_M k_P$. Thus, in this case, it is easy to see that changes in one parameter can be perfectly compensated by reciprocal changes in another parameter, such that unique values for these parameters cannot be fixed from the available observations.

To summarize, by the identifiability test of Section 3.1, we showed that the observation of reporter protein mean and variance profiles guarantees structural identifiability of the gene expression model, whereas the same model is not structurally identifiable from the mean only. This is similar to an analogous result in [8], where, however, global structural identifiability of a simpler gene expression model not accounting for switching promoter dynamics is investigated. Note that, for our case study, assessing global identifiability is rather challenging. For instance, by the existing approaches based on (13) [1,31], one would be confronted with an algebraic analysis of $Y(s,\theta)$, whose second entry is a ratio of polynomials of degree 10. Our local identifiability result was instead obtained by a symbolic computation approach that is fast, robust and fully general, as it can equally deal with any network in the class we consider.

For the same system, we now move on to a practical analysis of identifiability by the tools reported in Section 3.2. For simplicity, we fixed $z_0 = 0$ and assumed it known. We set the value of the rate parameters to $\theta^* = (k_M, d_M, k_P, d_P, \lambda_+, \lambda_-) = (0.5, 0.1, 0.2, 0.01, 0.05, 0.05)$ (in minutes^{-1}). These values are consistent with the literature [36,37] and correspond to a case of slow switching of the gene expression state and a stable reporter protein. (At this θ^*, the system is locally identifiable.) For convenience, let us denote by μ_P and σ_P^2 respectively the mean and variance of X_2 (the abundance of the reporter protein P), and use similar notation for the other species F and M. Simulated data

were generated for this system for $t \in [0, 95]$ (minutes). Random simulations were performed in StochKit [38]. We fixed measurement times to $t_\ell = (\ell - 1) \cdot 5$ min, with $\ell = 1, \ldots, N_y$ and $N_y = 20$. At these times, we computed empirical mean and variance statistics of $X(t)$ from 10^4 simulated trajectories (different simulations were used at different measurement times to mimic statistically independent cell samples across ℓ). This allowed us to to generate measurements of μ_P and σ_P^2 of the type (8) and (9). The obtained data is exemplified in Figure 2b. Based on 10 such datasets, we estimated θ^* 10 times by numerical solution of (16). Estimation was performed in MATLAB (`fmincon`, with initial guess $5 \cdot \theta^*$). Results are reported in Figure 2a together with 95% confidence regions determined from (18).

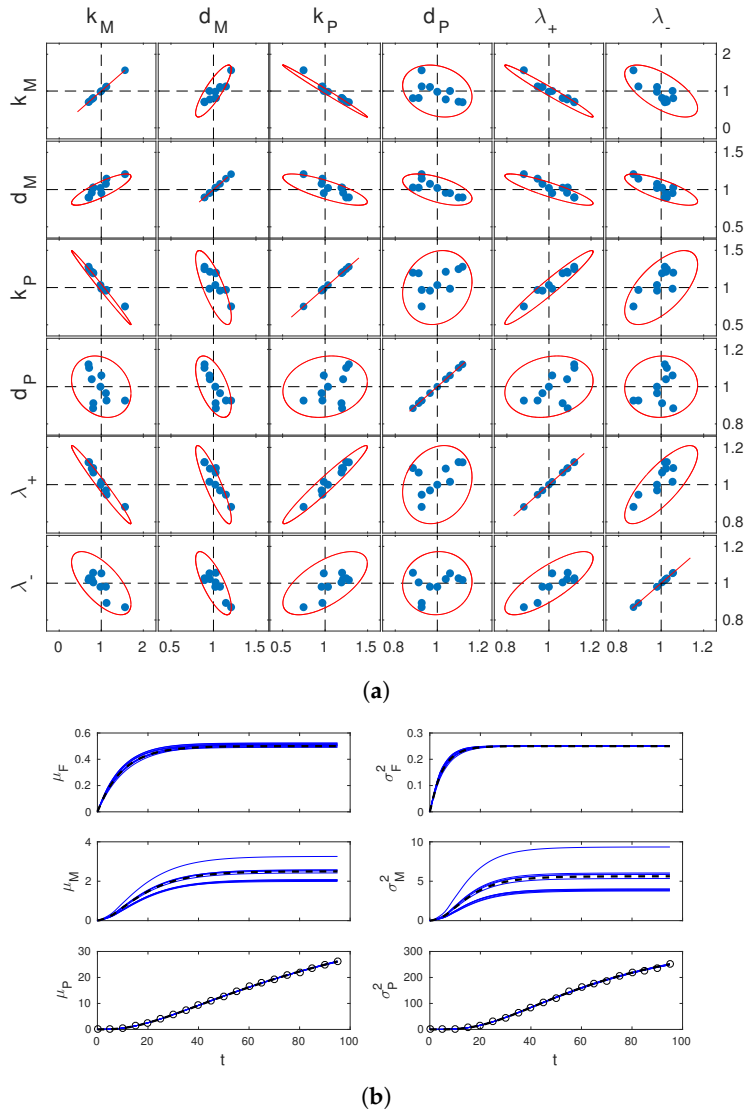


Figure 2. Parameter estimation results. (**a**) scatter plots of the estimates of parameters $(k_M, d_M, k_P, d_P, \lambda_+, \lambda_-)$ from 10 simulated datasets (blue dots) and theoretically computed 95% confidence regions (red lines). Results for all different pairs of these parameters are reported in the different boxes, as per labeling on top and left of the figure. Estimated values and pairwise confidence ellipsoids (one-dimensional confidence intervals for boxes on the diagonal) are normalized by the true parameter values. Dashed lines show the reference coordinates $(1, 1)$ corresponding to true values; (**b**) estimated dynamics of the system means (left) and variances (right), corresponding to the 10 different estimates of θ^* in (a). Solid blue lines show estimates, dashed black lines show true system statistics. In the bottom plots, black circles show measurements used for one of these estimates.

First of all, it can be seen that the confidence regions deduced from the approximate estimation error covariance matrix V are in excellent agreement wth simulated estimates. Therefore, they provide an effective tool to study practical identifiability. In our case, one sees from the figure that the confidence regions for the joint estimation of k_M and k_P are nearly degenerate ellipses. Relating this with the structural identifiability analysis carried out before, one concludes that, whereas variance observations guarantee identifiability of the system, discerning k_M and k_P unambiguously from their product $k_M k_P$ remains difficult. Further analysis of Figure 2a shows that discerning λ_+ from k_M and k_P is also difficult. This is not surprising since λ_+ itself multiplies k_M and k_P in the entry of $Y(s,\theta)$ displayed above (however λ_+ also appears at the denominator separate from the other parameters). Finally, it is interesting to look at the time profiles of the statistics of M and F corresponding to the 10 estimates of θ^*. These are shown in Figure 2b. The estimates of $\mu_F(\cdot)$ and $\sigma_F^2(\cdot)$ are all close to the true profiles, while those of $\mu_M(\cdot)$ and $\sigma_M^2(\cdot)$ are significantly more dispersed. Apparently, the limited accuracy in the estimation of k_M, k_P and λ_+ is reflected in the estimates of the mRNA statistics, but not in those of gene activation statistics. This shows that, depending on the purpose of parameter identification, inaccurate parameter estimates may or not be detrimental for the study of the system.

In summary, the study of practical identifiability of the reporter gene system showed that structural identifiability translates into the ability to discern all parameters unambiguously, but with an accuracy that still depends on the interplay of the parameters. Moreover, certain system dynamics may still be predicted with accuracy in spite of poor estimates of relevant system parameters. This relates with the concept of "parameter sloppiness" discussed in the literature for biological systems [5].

4. Identification of Networks

As discussed in Section 2, the structure of regulatory interactions among species of a given network is captured by the pattern of nonzero entries of S and W. In Section 3, we have discussed the problem of identifying unknown parameters of a reaction network with a known structure of interactions. By the case study of Section 3.3, we have shown that usage of stochastic information, such as the dynamics of the network state variance, may guarantee structural parameter identifiability in cases where the sole usage of deterministic information, such as the mean system response, does not. In this section, we wish to investigate whether the addition of stochastic information also helps identification of networks with unknown interactions. In full generality, the problem thus becomes the identification of the matrices $S \in \mathbb{Z}^{n\times m}$ and $W \in \mathbb{R}^{m\times n}$, defining models (10) and (11) via (5), from measurements of type (8). Matrix C, describing what statistical moments are observed for what species, is fixed by the measurement experiment design and is thus known.

In a first, naive attempt, one could approach identification of S and W by directly fitting measurements (8) with the predictions of (10) and (11). However, this is computationally intensive, since it requires solving the dynamics (10) repeatedly over the joint search space of S and W, and does not give any insight into the identifiability of S and W. On the other hand, Equations (10) and (11) are a linear dynamical system. Disregarding the specific structure of A and B, linear dynamical systems can be reconstructed by standard identification techniques [31,39]. This suggests separating identification of S and W into two steps. The first step is the reconstruction of a dynamical model in the form

$$\begin{aligned} \dot{\xi}(t) &= \hat{A}\xi(t) + \hat{K}u(t), \\ y(t) &= \hat{C}\xi(t), \end{aligned} \tag{22}$$

from measurements (8). For noiseless measurements, one such model perfectly matching the data is of course given by Equations (10) and (11), with $z = \xi$, $\hat{A} = A$, $\hat{K} = BG$ and $\hat{C} = C$. Assuming for a moment that the reconstructed $\hat{A}$ equals A and $\hat{K} = BG$, the second step is the inference of S and W from $\hat{A}$ and $\hat{K}$ in light of Equation (5). A similar approach is used for parameter estimation of a known network in [30].

A priori, none of the two steps has a unique solution. In the first step, several models $(\hat{A}, \hat{K}, \hat{C})$ may equally explain the same data, that is, $\hat{A}$ and $\hat{K}$ may not be uniquely defined (even their dimension is generally not uniquely determined). In addition, noise on measurement samples propagates into the computation of $\hat{A}$, $\hat{K}$. In the second step, leaving G aside for the moment, the crucial question is whether the matrices S and W that define the network structure can be uniquely factored out of $\hat{A}$. As we will see, the answer depends in the first place on the relationship between $\hat{A}$ and A.

In analogy with the discussion of Section 3.1, we will first look at the problem in terms of identifiability from $y(\cdot)$ and $u(\cdot)$. In Section 4.1, we will look at the reconstruction problem of a suitable model in the form (22). In Section 4.2, we will look at the subsequent problem of extracting S and W from the reconstructed $\hat{A}$. In order to understand how stochasticity contributes to the reconstruction of unknown networks, in these two sections, we will focus on a first case, where $y(\cdot) = \mu(\cdot)$, and a second case, where observations additionally include the entries of $\Sigma(\cdot)$, that is, $y(\cdot) = z(\cdot)$. Then, in Section 4.3, we will look at these two steps from a practical perspective, providing a method to infer S and W from real data. The contribution of stochasticity to the identifiability of S and W and the full network identification procedure will be demonstrated on an example in Section 4.4. We will assume throughout this section that $u(t)$ is assigned, leaving experiment design considerations to later studies (see Section 5).

4.1. Step 1: Identifiability of a Linear Model for the Moment Dynamics

Here, we are concerned with the reconstruction of a model in the form (22). Let us recall some facts [40]. In the context of linear state-space models, a realization of a (strictly causal) linear map $\mathscr{L} : \mathscr{U} \to \mathscr{Y}$ from an input space $\mathscr{U}$ to an output space $\mathscr{Y}$ is any triplet $(\hat{A}, \hat{K}, \hat{C})$ such that the solution of (22) reproduces this mapping. Its order is the size of matrix $\hat{A}$. It is well known that, if $(\hat{A}, \hat{K}, \hat{C})$ is one realization, all triplets $(T\hat{A}T^{-1}, T\hat{K}, \hat{C}T^{-1})$, with T a nonsingular square matrix, are equally viable realizations. A realization is minimal if it has the smallest size of $\hat{A}$ among all realizations. If $(\hat{A}, \hat{K}, \hat{C})$ is minimal, then all minimal realizations are of the form $(T\hat{A}T^{-1}, T\hat{K}, \hat{C}T^{-1})$ for some nonsingular T, i.e., minimal realizations form an equivalence class.

Recall that n_y is the number of rows of C, that is, the size of vector y. We will make the following assumption.

Assumption 1. *The minimal realization of the linear map expressed by* (10) *and* (11) *is of order greater than or equal to n_y. In addition, for the given u, the class of minimal realizations is uniquely determined by u and the corresponding y.*

The first part of this assumption concerns structural properties of the system, essentially ruling out singular cases where the moment equations provide a redundant description of the dynamics of y. As a counterexample, a system that would not fulfill this requirement is one where two observed species participate in the same reaction in exactly the same way, such that the evolution of their moments is identical. On the basis of the first part, the second part of the assumption concerns the informativity of the given input–output pair, requiring in practice that all observed moments are excited by the given input. In view of the relation between minimal realizations and transfer functions, this corresponds to saying that the system transfer function can be identified from the given input–output pair. This is a reasonable requirement provided a sufficiently rich excitation input [39,40].

In summary, under Assumption 1, identification is well-posed up to an unknown matrix T. However, if $(\hat{A}, \hat{K}, \hat{C})$ is a minimal realization determined from u and y, T is constrained by the knowledge of C, since it must hold that $\hat{C}T^{-1} = C$. What this constraint implies on the reconstructed model depends on how a minimal realization of the system itself looks like. We focus on the following cases:

Case (i): Observation of mean only ($y = \mu$). In this case, $n_y = n$ and $C = \begin{bmatrix} C' & 0_{n\times n^2} \end{bmatrix}$, with $C' \in \mathbb{R}^{n\times n}$ nonsingular (typically the identity). In view of the structure of A in (5), for this definition of C, one realization of (10) and (11) is

$$\begin{aligned} \dot{\xi}(t) &= SW\xi(t) + SGu(t), \\ y(t) &= C'\mu(t). \end{aligned}$$

This realization is of order $n_y = n$ and is minimal for non-degenerate definitions of S, W and G. Assumption 1 is thus satisfied provided the input and/or the initial conditions excite all system dynamics. Then, any reconstructed model $(\hat{A}, \hat{K}, \hat{C})$ must satisfy $(\hat{A}, \hat{K}, \hat{C}) = (TSWT^{-1}, TSG, C'T^{-1})$ for some invertible T. Since C' is known and invertible, $T = \hat{C}^{-1}C'$ is uniquely determined, and so are $SW = T^{-1}\hat{A}T$ and $SG = T^{-1}\hat{K}$.

Case (ii): Observation of mean and covariance matrix ($y = z$). Since $\Sigma = \Sigma^T$, this case is captured by a model where C has $n_y = n + n(n+1)/2$ rows and $n + n^2$ columns. The definition of C is such that $y = Cz = C''\xi$, where ξ is an n_y-dimensional vector containing all and only the distinct entries of z, and C'' is invertible (in particular, C and C'' can be $(0,1)$-matrices). One realization of (10) and (11) is then

$$\begin{aligned} \dot{\xi}(t) &= A''\xi(t) + K''u(t), \\ y(t) &= C''\xi''(t), \end{aligned}$$

where A'' and K'' are formed from all and only the distinct entries of A and BG, in the same order, in accordance with the definition of ξ. This realization is of order n_y, and is minimal except for peculiar definitions of S and W. Similar to Case (i), provided Assumption 1 is satisfied, any reconstructed model $(\hat{A}, \hat{K}, \hat{C})$ must satisfy $(\hat{A}, \hat{K}, \hat{C}) = (TA''T^{-1}, TK'', C''T^{-1})$, with T invertible. Because C'' is known, by the same arguments used in Case (i), matrices A'' and K'' are uniquely determined. Since they necessarily contain all elements of A and BG, the latter are also uniquely determined.

In summary, under Assumption 1, Case (i) guarantees unambiguous reconstruction of the products SW and SG, while Case (ii) also guarantees unambiguous reconstruction of the other blocks of A, namely $S^{(2)}W$ and $I_n \otimes (SW) + (SW) \otimes I_n$, and of the second block of BG, namely $S^{(2)}G$. Other cases of interest exist, among which, for instance, the observation of the statistics of part of X, or the observation of the diagonal entries of Σ only [8,9,18]. The investigation of these scenarios is beyond the scope of this paper, but is rediscussed in Section 5.

4.2. Step 2: Identifiability of the Network Stoichiometry and Rate Parameter Matrices

We have shown in the previous section that different information about the matrices of model (10) and (11) can be obtained depending on the definition of matrix C. This can be summarized in terms of the $(n + n^2) \times (n + n_u)$ matrix

$$\Xi(S, W, G) \triangleq \begin{bmatrix} SW & SG \\ S^{(2)}W & S^{(2)}G \end{bmatrix}. \tag{23}$$

Under Assumption 1, Case (i) yields perfect reconstruction of the first block-row (first n rows) of $\Xi(S, W, G)$, whereas Case (ii) also yields the remaining n^2 rows. Note that, for the latter case, the further availability of $I_n \otimes (SW) + (SW) \otimes I_n$ does not provide additional information since the product SW is already part of the first n rows of $\Xi(S, W, G)$. For these two cases, the question addressed in this section is what can be said about the individual contributions of S, W and G.

Denote with $\Xi_h(S,W,G)$ the first h rows of (23). For $h=n$ and $h=n+n^2$, in the order, denote with $\hat{\Xi}_h$ the estimate of $\Xi_h(S,W,G)$ obtained in Step 1 for Case (i) and Case (ii), in the same order. Under the current hypothesis of perfect reconstruction, it holds that

$$\hat{\Xi}_h = \Xi_h(S,W,G). \tag{24}$$

The question we are posing is about the solutions (S,W,G) of (24) given $\hat{\Xi}_h$. One solution to (24) is of course provided by the true network matrices, which we denote S^*, W^*, G^*. Equivalent solutions are all triplets (S,W,G) obtained by permutation of the columns of S^* and corresponding permutation of the rows of W^* and G^*, since these are trivially different enumerations of the network reactions $\mathscr{R}_1,\ldots,\mathscr{R}_m$. However, other solutions may exist. Simple algebraic manipulations of (23) and (24) yield

$$\hat{\Xi}_h = L_h(S)\cdot\begin{bmatrix} W & G \end{bmatrix}, \qquad L_h(S) \triangleq \text{first } h \text{ rows of } \begin{bmatrix} S \\ S^{(2)} \end{bmatrix}. \tag{25}$$

Thus, for a given $\hat{\Xi}_h$, the viable triplets (S,W,G) are all solutions to the mixed factorization problem (25) with discrete-valued factor $L_h(S)$ and continuous-valued factor $[W\ G]$. Clearly, the set of solutions is generally smaller the larger the h, since more equations need to be satisfied. That is, Case (ii) has better potential than Case (i) for precise network reconstruction. Starting from (25), a characterization of the solutions in terms of algebraic or topological properties of the true underlying network would be desirable but is still unavailable to us. Our contribution here is an operational definition of all possible solutions, that is, an algorithmic procedure to seek all solutions corresponding to a given $\hat{\Xi}_h$.

Let $\mathbb{S}\subseteq\mathbb{Z}^{n\times m}$, $\mathbb{W}\subseteq\mathbb{R}^{m\times n}$ and $\mathbb{G}\subseteq\mathbb{R}^{m\times n_u}$, with $\mathbb{W}$ and $\mathbb{G}$ convex. For any matrix norm $|||\cdot|||$, if $S^*\in\mathbb{S}$, $W^*\in\mathbb{W}$ and $G^*\in\mathbb{G}$, the solutions of (25) coincide with the set of optimizers of

$$Q^* \triangleq \min_{(S,W,G)\in\mathbb{S}\times\mathbb{W}\times\mathbb{G}} Q(S,W,G), \qquad Q(S,W,G) \triangleq |||\hat{\Xi}_h - L_h(S)\cdot[W\ G]|||, \tag{26}$$

and (S^*,W^*,G^*) attains the minimum $Q^*=0$. Problem (26) can be rewritten as

$$Q^* = \min_{S\in\mathbb{S}} \hat{Q}(S), \tag{27}$$

$$\hat{Q}(S) \triangleq \min_{(W,G)\in\mathbb{W}\times\mathbb{G}} Q(S,W,G). \tag{28}$$

For any fixed S, Equation (28) is a least-squares problem with convex constraints that is easy to solve by standard search algorithms [41]. For finite $\mathbb{S}$, the complete solution space of (27) can then be determined by exploration of $\mathbb{S}$, seeking all S such that $\hat{Q}(S)=0$. In general, the solution of (28) is not unique. If S is such that $\hat{Q}(S)=0$ and (W,G) solves the corresponding problem (28), then all couples $(W+\tilde{W},G+\tilde{G})\in\mathbb{W}\times\mathbb{G}$ such that the columns of $[\tilde{W}\ \tilde{G}]$ are in $\ker\big(L_h(S)\big)$ ("$\ker(\cdot)$" denotes the kernel of a matrix) equally solve (28). The multiplicity of the solutions for a given S thus depends on the interplay between $\ker\big(L_h(S)\big)$ and $\mathbb{W}\times\mathbb{G}$. We will come back on this in the example of Section 4.4.

For all practical matters (robustness to numerical errors and modelling inaccuracies, and applications below where $\hat{\Xi}_h$ is computed from the noisy data (8), one should require that the equality in (25) holds approximately within a suitable tolerance $\epsilon>0$. This leads to the procedure for the computation of the set of solutions Ω detailed in Algorithm 1.

One possible choice of $\mathbb{S}$ is to consider all matrices $S\in\mathbb{Z}^{n\times m}$ whose elements $S_{i,j}$ are such that, for some $S_{max}\in\mathbb{N}$, $|S_{i,j}|\leq S_{max}$, with $i=1,\ldots,n$ and $j=1,\ldots,m$. The size of $\mathbb{S}$ in this case is of order $\mathscr{O}(S_{max}^{nm})$. Despite the exponentially growing complexity, exploring $\mathbb{S}$ by enumeration remains viable for networks of small size. The complexity of this search could be dramatically reduced by recalling that permutations of the reactions list amount to equivalent models. Other ameliorations are possible to improve the scalability of the method (see also the relevant discussion in Section 5).

Algorithm 1: Identification of stoichiometry and rate parameters from a model of the moment dynamics

Given $\hat{\Xi}_h$ and an $\epsilon > 0$:
Set $\Omega = \emptyset$;
For every $S \in \mathbb{S}$:
  Solve problem (28) to get $\hat{Q}(S)$ and the solution set $\hat{\Omega}(S) = \{(W, G) : \ Q(S, W, G) = \hat{Q}(S)\}$;
  If $\hat{Q}(S) < \epsilon$, include $\{S\} \times \hat{\Omega}(S)$ in Ω;
Return Ω.

To conclude, suppose that the true number of reactions, say m^*, is unknown. One way to generalize the procedure above to this scenario is to execute it for a value of m large enough. By this approach, let $\hat{m} \leq m$ be the minimum number of nonzero columns of S among the elements of Ω. Since the null columns of S do not contribute to the network dynamics, $\hat{m}$ is the minimum number of reactions needed to explain the data and is thus a viable estimate of m. In practice, though, this approach is computationally inefficient, since it explores an unnecessarily large space $\mathbb{S}$. A better solution is to proceed incrementally and execute the algorithm for increasing values of m, stopping the exploration as soon as $\Omega \neq \emptyset$.

4.3. Network Identification in Practice

So far in this section, we have treated reconstruction of moment dynamics (Step 1) and then of biochemical network matrices (Step 2) in the absence of noise. We now devise a procedure for the two-step estimation of S, W and G from noisy moment measurements of the types (8) and (9). To do this, we will repeatedly exploit the methods of Section 3.2, with a specific definition of parameters θ. For ease of exposition, we focus on Case (ii) and assume that the initial system moments z_0 are known. The generalization to possible unknown entries of z_0 is straightforward. The necessary adaptations for Case (i) are commented on at the end of the section.

Consider Step 1 first. As mentioned in Section 4.1, general approaches to the estimation of linear state-space models can be borrowed from literature. Here, however, we account explicitly for the structure of model (5) and develop a dedicated approach. Let $\theta_{1,1}$, $\theta_{1,2}$, $\theta_{2,1}$ and $\theta_{2,2}$ denote the unknown matrix products SW, SG, $S^{(2)}W$ and $S^{(2)}G$ that compose (23), in the same order, and let θ be the vector collecting all entries of $\theta_{1,1}$, $\theta_{1,2}$, $\theta_{2,1}$ and $\theta_{2,2}$. Define $\hat{y}_\theta(\cdot)$ as the solution of

$$\frac{d}{dt}\hat{z}_\theta = \begin{bmatrix} \theta_{1,1} & 0 \\ \theta_{2,1} & I_n \otimes \theta_{1,1} + \theta_{1,1} \otimes I_n \end{bmatrix} \hat{z}_\theta + \begin{bmatrix} \theta_{1,2} \\ \theta_{2,2} \end{bmatrix} u(t), \qquad \hat{z}_\theta(0) = z_0, \tag{29}$$

$$\hat{y}_\theta = C\hat{z}_\theta. \tag{30}$$

Note that, for true parameter values, these equations coincide with the true moment dynamics (10) and (11). With these definitions of θ and $\hat{y}_\theta$, for a suitable search space Θ, we compute the ML estimate $\hat{\theta}$ by the solution of (16), which provides us with estimates of SW, SG, $S^{(2)}W$ and $S^{(2)}G$. In practice, this solution shall be computed by numerical optimization.

Now, consider Step 2. The estimates $\hat{\theta}$ of SW, SG, $S^{(2)}W$ and $S^{(2)}G$ from above allow us to build (a noisy version of) the matrix $\hat{\Xi}_h$ of Section 4.2 (with $h = n + n^2$ for Case (ii)). This matrix can be used to run Algorithm 1 and determine a set of solution triplets (S, W, G). However, for a candidate solution (S, W, G), the acceptance criterion $Q(S, W, G) < \epsilon$ (with $Q(S, W, G) = |||\hat{\Xi}_h - L_h(S) \cdot [W, G]|||$) needs to be adapted to the statistics of the noise that corrupts $\hat{\Xi}_h$. To do this, we first compute the estimation error covariance matrix for $\hat{\Xi}_h$, say V_h, by the sensitivity method explained after (16). Armed with V_h, for a given confidence level $1 - \alpha$, the idea is to define the norm $|||\cdot|||$ and ϵ such that a candidate

solution (S, W, G) is accepted by Algorithm 1 if the discrepancy between $L_h(S) \cdot [W\ G]$ and $\hat{\Xi}_h$ falls within the $(1-\alpha)$-confidence ellipsoid around $\hat{\Xi}_h$. In view of (18), this is obtained by setting $\epsilon = \chi_\alpha$ and

$$Q(S, W, G) = \mathrm{vec}\big(\hat{\Xi}_h - L_h(S) \cdot [W\ G]\big)^T V_h^{-1} \mathrm{vec}\big(\hat{\Xi}_h - L_h(S) \cdot [W\ G]\big).$$

In this way, our acceptance criterion is expressed as a χ^2-statistical test that the candidate solution corresponds to the true parameters underlying our estimate $\hat{\Xi}_h$.

For Case (i), the method described above remains the same, except that θ shall only contain the entries of $\theta_{1,1}$ and $\theta_{1,2}$, model (29) and (30) is restricted to the mean dynamics, and $\hat{\Xi}_h$, L_h are defined for $h = n$. From a computational viewpoint, our two-step approach requires fitting moment equations to measurements (8) only once (Step 1). On the contrary, the naive approach discussed at the beginning of Section 4 would require solving a similar optimization problem iteratively for all the candidate stoichiometries S. Our approach postpones this search to the second step based on the iterative solution of a much simpler, convex optimization problem, with tremendous computational saving. Observe that the error statistics in the estimation of matrix products S^*W^* and $(S^*)^{(2)}W^*$ from Step 1 are quantitatively accounted for in the construction of the test to select the compatible network structures in Step 2. Therefore, provided the ML estimator in Step 1 is close to optimality, the splitting of network identification into two steps is not expected to deteriorate performance relative to a one-step procedure testing candidate network structures directly on the moment measurements.

In summary, we have devised an algorithm to perform reconstruction of biochemical networks from real-world population-snapshot data. The method can be applied to mean data only, but, contrary to existing approaches, it equally applies to joint mean and variance measurements. We argued by theoretical arguments that leveraging the additional variance measurements is expected to improve reconstruction accuracy. In the next section, we will show that this is indeed the case by the analysis of a case study, which reconfirms the practical interest of our method.

4.4. Example: A Toy Network

We now apply the methods developed above to a toy example. Our first aim is to study the achievable network reconstruction performance in Case (i) (observations of mean only) and (ii) (observations of mean and covariance matrix). To do this, we initially focus on the network reconstruction step described in Section 4.2, assuming that the matrix products S^*W^* and, for Case (ii), $(S^*)^{(2)}W^*$ are known exactly. For simplicity, we ignore G. This constitutes no loss of generality since G enters the factorization problem (25) in a way similar to W. Our second aim is to test the full identification procedure of Section 4.3, where the products S^*W^* and $(S^*)^{(2)}W^*$ are not known and need to be estimated from noisy measurements of the network moments. This is done at the end of the section on the basis of simulated data.

Consider a reaction network with

$$S^* = \begin{bmatrix} -1 & 0 & 0 \\ 1 & 1 & -1 \end{bmatrix}, \quad W^* = \begin{bmatrix} 0.2 & 0 \\ 0.1 & 0 \\ 0 & 0.1 \end{bmatrix},$$

where superscript "$*$" denotes true system quantities. The actual network size is then $n^* = 2$ and $m^* = 3$. For this system, the matrix products defining the moment dynamics are

$$S^*W^* = \begin{bmatrix} -0.2 & 0 \\ 0.3 & -0.1 \end{bmatrix}, \qquad (S^*)^{(2)}W^* = \begin{bmatrix} 0.2 & 0 \\ -0.2 & 0 \\ -0.2 & 0 \\ 0.3 & 0.1 \end{bmatrix}.$$

First assume that, under Assumption 1, these products are perfectly reconstructed from the observation of the system moments. More precisely, in agreement with Section 4.1, S^*W^* is available for Case (i) and (ii), whereas $(S^*)^{(2)}W^*$ is additionally available for Case (ii) only. In order to infer S^* and W^* from this data by the procedure of Section 4.2, we assumed that $\mathbb{S}$ is the set of all integer matrices with entries $|S_{i,j}| \leq S_{max}$, with $S_{max} = 2$, and $\mathbb{W} = \mathbb{R}_+^{m\times n}$. We first tested the approach with the number of reaction channels known and equal to the true value $m = m^* = 3$. The size of $\mathbb{S}$ in this case is $(2S_{max}+1)^{n\cdot m} = 5^6 = 15,625$ possible stoichiometry matrices. For every candidate stoichiometry S, the computation of $\hat{Q}(S)$ and $\hat{\Omega}$ was performed via the Matlab function `lsqnonneg`, implementing quadratic optimization under nonnegativity constraints. To cope with numerical errors in the solution of this optimization, here we take $\epsilon = 10^{-6}$. Results are summarized in Table 1 (column with $m = 3$).

Table 1. Network reconstruction results for Case (i) and Case (ii), for different hypotheses on the number of reactions m. Number of solutions refers to the number of different stoichiometry matrices in Ω. Acceptance ratio is the number of solutions divided by the number of stoichiometry matrices tested, given by $|\mathbb{S}| = 5^{2m}$. Computational times are in seconds, evaluated on a 4-core 3GHz Intel Xeon processor (Santa Clara, CA, USA). Results for the true number of reactions ($m = m^*$) are reported in bold.

	m	1	2	3	4
Case (i)	Number of solutions	0	4	**2604**	150,172
	Acceptance ratio	0%	0.64%	**16.7%**	38.4%
	Computational time	<0.01	0.11	**2.86**	75.13
Case (ii)	Number of solutions	0	0	**6**	564
	Acceptance ratio	0%	0%	**0.038%**	0.14%
	Computational time	0.01	0.12	**3.11**	80.57

In Case (i), we found 2604 stoichiometry matrices S such that $\hat{Q}(S) < \epsilon$, i.e., about 16.7% of the matrices in $\mathbb{S}$ can explain the data, provided an adapted choice of rate parameters W. For the given definition of $\mathbb{W}$ (nonnegative real vectors), if (S, W) is a solution and the kernel of S contains a nonnegative vector $\tilde{W} \in \mathbb{W}$, then all couples $(S, W + c\tilde{W})$ with $c \geq 0$ are solutions, since $c \cdot \tilde{W} \in \mathbb{W}$ and $S \cdot (W + c \cdot \tilde{W}) = SW$ (note that the kernel of S is nontrivial for $m > 2$). Thus, in general, infinitely many solutions (S, W) correspond to a viable S.

In Case (ii), only six matrices S were found such that $\hat{Q}(S) < \epsilon$, i.e., only about 0.038% of all possible stoichiometries $\mathbb{S}$ is in agreement with the data. This is of course due to the different definition of Q, which also penalizes poor fit with $(S^*)^{(2)}W^*$. In this case, W is uniquely determined by S, since $\tilde{W} \in \mathbb{W}$ should be sought in $\ker(S) \cap \ker(S^{(2)})$, and all S that were selected are such that $\ker(S) \cap \ker(S^{(2)}) = \{0\}$. Compared with Case (i) above, the gain in using observations of the second-order moments is thus striking.

In both cases, as anticipated in Section 4.2, the solutions found are redundant, since they correspond to the same reactions listed in a different order. In particular, we verified by inspection that the six solutions (S, W) in Case (ii) are given by all possible column permutations of S^* and corresponding permutations of the rows of W^*. Thus, in this case, the solution found is essentially unique. In Case (i), these 6 solutions are instead contained in a much larger pool of 2604 putative solutions.

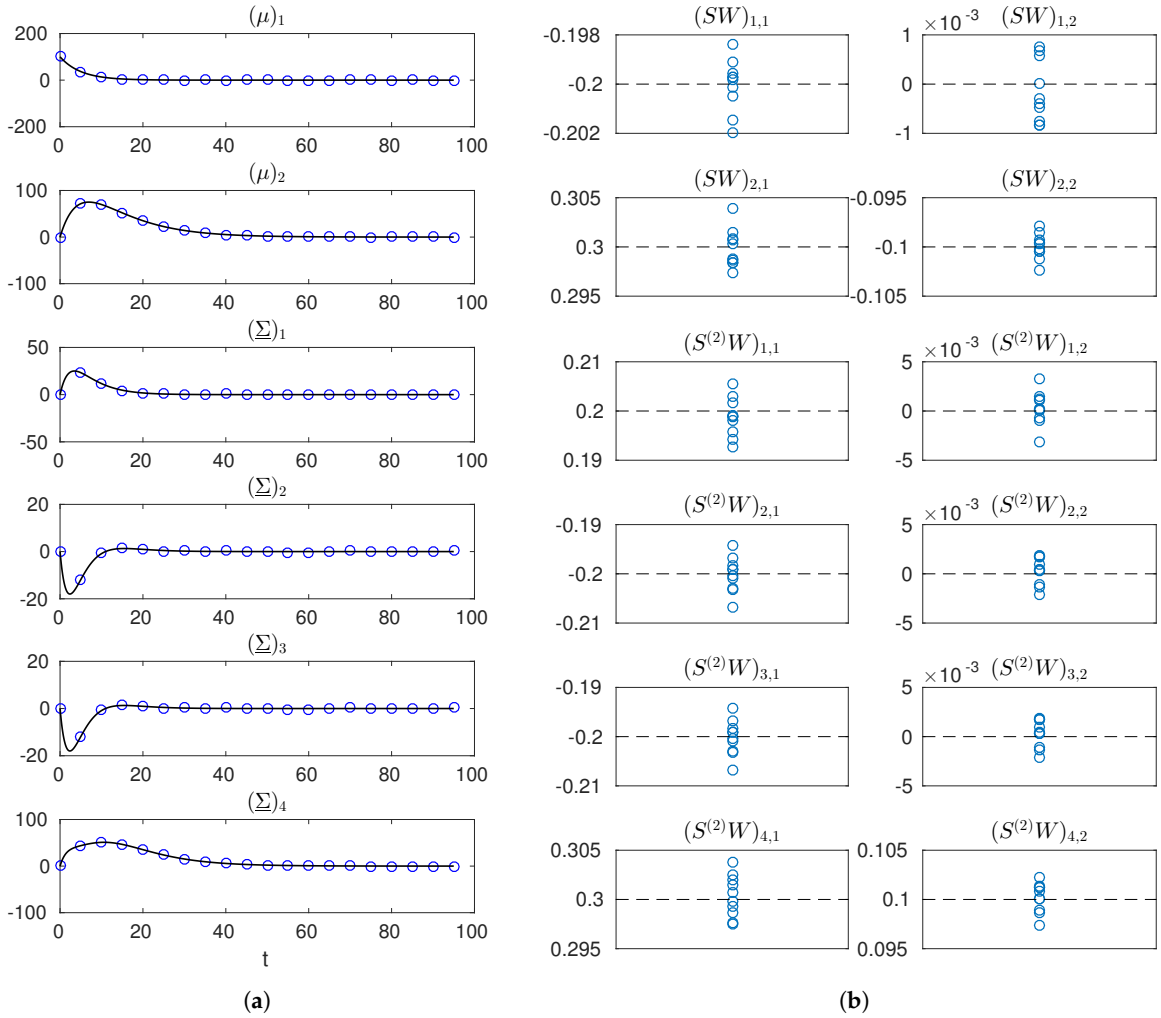
We then run the algorithm for values of m different from m^* in order to test the feasibility of its estimation. In Case (i), a nonempty solution set Ω was returned for $m \geq 2$. This is not surprising since the columns of S^*W^* belong to a two-dimensional linear space. Therefore, in general, S must have at least two columns in order to span this space via the product SW. At the same time, this shows that m^* will be underestimated based on sole mean measurements. In Case (ii), instead, solutions with $m < m^*$ were ruled out, showing the potential of joint mean and variance measurements for

the estimation of m^*. In both Case (i) and (ii), for $m = 4$, many putative solutions are found from within the pool of $5^8 = 390,625$ stoichiometries tested. Putative solutions include correct solutions where S has one null column and the remaining columns given by a permutation of the columns of S^*. As expected, alternative pairs (S, W) that do not relate with (S^*, W^*) as also returned in Ω, as a result of the over-parametrization of the model in this case. However, Table 1 confirms that the covariance measurements collected in Case (ii) guarantee a tighter selection of the solution set.

The computational time to run the algorithm increases rapidly with m, as expected from the exponential complexity of the search, and it is similar for Case (i) and (ii). In scenarios where m^* is unknown, this shows the importance of testing candidate values m incrementally. In the given example, stopping the search with the solutions found for $m = 3$ (overall execution time less than four seconds) allows one to spare much computational time associated with the exploration of solutions with $m = 4$ (between one and two minutes).

Finally, for Case (ii), we implemented and run the full identification procedure of Section 4.3 from simulated data. Here, the products S^*W^* and $(S^*)^{(2)}W^*$ are not known and are estimated from noisy mean and variance measurements. These measurements were obtained by simulating the moment Equations (10) and (11) with the true S^* and W^* from initial conditions known and fixed to $\mu(0) = [100, 0]^T$ and $\Sigma(0)$ null. We generated measurements at times $t_\ell = (\ell - 1) \cdot T$, with $\ell = 1, \ldots, N_y$, where $T = 5$ and $N_y = 20$, adding random noise with covariance matrix $R_\ell = \mathrm{diag}(0.3^2, 0.3^2)$ for all ℓ, corresponding to roughly 1% error on the observed mean and variance profiles. An example simulated dataset is shown in Figure 3a. The numerical optimization in the first step of the procedure, yielding noisy estimates of the matrix products S^*W^* and $(S^*)^{(2)}W^*$, was implemented by MATLAB's `lsqnonlin`. Results are illustrated in Figure 3b. Estimates are found to be well-centered and little dispersed relative to the true values of the entries of S^*W^* and $(S^*)^{(2)}W^*$, showing the effectiveness of the reconstruction of the moment dynamics. Starting from the noisy estimates of S^*W^* and $(S^*)^{(2)}W^*$, the second step was implemented as described in Section 4.3, for a significance level of $\alpha = 0.05\%$. With this noise and significance levels, for $m \in \{1, 2, 3, 4\}$, the same solutions as in Table 1 were returned over several runs, showing feasibility and effectiveness of the approach. For the case of $m = m^*$, in particular, we quantified the rate of rejection of correct candidate solutions. Over a few hundred runs, we found this rate to be around 1%, that is, smaller than the prescribed rate α. This can be ascribed to the linear approximations made to establish the χ^2 statistic in Section 3.2. On the other hand, incorrect solutions were never accepted.

To sum up, we showed that stochastic information about the network dynamics allows one to identify the structure of a biochemical reaction network in cases where the sole mean data does not. In addition, in the scenario where the observable moments provide full information for network reconstruction, we showed that our two-step procedure is capable of correctly identifying the example network from data with realistic measurement error levels.



(**a**) (**b**)

Figure 3. Simulated measurements and estimates of S^*W^* and $(S^*)^{(2)}W^*$ for the network reconstruction example. (**a**) true trajectories of the entries of μ and $\underline{\Sigma}$ (solid black line) and one simulated dataset (blue markers); (**b**) estimates of the entries of S^*W^* and $(S^*)^{(2)}W^*$ (blue markers) obtained from 10 different datasets (dashed black lines indicate true values). Notation $(\cdot)_{r,c}$ is used in labels to denote the row-r, column-c entry of a matrix.

5. Discussion

In this paper, we have investigated the problems of parameter identification and reconstruction of interactions in biochemical reaction networks. In the context of population snapshot data (first- and second-order statistics), and state-affine reaction rates, we have provided practical methods to study identifiability of unknown parameters and algorithms to address network reconstruction. In both cases, we have shown superiority of stochastic, single-cell approaches over deterministic, population-average data. In so doing, we have also extended the existing results on identifiability of gene expression models.

Our parameter identifiability analysis is developed with reference to a given observation model (what statistics of what species are detected) and a given perturbation input. A desirable theoretical advance of our work is the investigation of identifiability as a property independent of a specific input choice. From a practical standpoint, instead, evaluating structural and practical identifiability for different inputs and observation models allows one to design most informative experiments and, for synthetic biology, the engineering of most informative intracellular reporters. Developments of

our work in the context of optimal experiment design indeed constitute a first important direction of future investigation.

Our results were developed for a class of problems of immediate relevance to applications. Population-snapshot data are nowaday easily collected by simple experiments such as flow-cytometry. More complex experimental setup, such as microfuidics in combination with video-microscopy, also provide this type of data. In addition, these allow for time-lapse monitoring of individual cells, providing time correlations that we did not account for here. However, this requires nontrivial processing of raw images entailing single-cell tracking procedures, which are not always performed in practice. Instead, for network reconstruction, more complex measurement scenarios, such as the monitoring of a subset of the network species and lack of covariance measurements across different species, shall be addressed in detail to widen the applicability of our methods.

Of course, the choice of networks with affine rates formally restricts one to reactions of zeroth- or first-order only. However, as the random-telegraph reporter gene model exemplifies, real systems of interest exist in this form. At the same time, state-affine rates have been used to model and investigate interactions of arbitrarily complex networks in an approximate manner [15]. Precise account of nonlinear stochastic dynamics instead requires generalization of our methods, for instance, with moment-closure approaches [29,42,43]. This is another direction of research.

Finally, the proposed network reconstruction methods were shown to be viable for a small toy example. Whereas automated reconstruction of small networks is of practical interest, network reconstruction methods are of greatest help for the investigation of larger networks of interactions. Our current algorithm performs fast on small networks but scales poorly with the network size (number of species and putative reaction channels). As anticipated in Section 4.2, great improvements can be easily obtained by an optimized implementation that avoids the exploration of redundant network candidates. In particular, considering that not only permutations of the reactions list but also identical columns of a stoichiometry matrix are redundant, it can be seen that the number of effectively distinct stoichiometry matrices to be tested is in the order of $\binom{S^n_{max}}{m}$. Therefore, an optimized implementation of the method would reduce complexity from $\mathscr{O}(S^{nm}_{max})$ to $\mathscr{O}\left(\binom{S^n_{max}}{m}\right)$. However, the discrete nature of the stoichiometry matrix makes the problem inevitably hard, and advanced mixed-integer programming methods should be explored in order to tackle the exponential growth of complexity with the network size. This is yet another research direction of practical relevance.

Funding: This work was funded in part by the French National Research Agency (ANR) via project MEMIP: Mixed-Effects Models of Intracellular Processes (ANR-16-CE33-0018), and by the Inria Project-Lab (IPL) CoSy.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Ashyraliyev, M.; Fomekong-Nanfack, Y.; Kaandorp, J.; Blom, J. Systems Biology: Parameter Estimation for Biochemical Models. *FEBS J.* **2009**, *276*, 886–902. [CrossRef] [PubMed]
2. Marbach, D.; Costello, J.; Küffner, R.; Vega, N.; Prill, R.; Camacho, D.; Allison, K.; The DREAM5 Consortium; Kellis, M.; Collins, J.; et al. Wisdom of crowds for robust gene network inference. *Nat. Methods* **2012**, *9*, 796–804. [CrossRef] [PubMed]
3. Purnick, P.; Weiss, R. The second wave of synthetic biology: From modules to systems. *Nat. Rev. Mol. Cell Biol.* **2009**, *10*, 410–422. [CrossRef] [PubMed]
4. Chis, O.T.; Banga, J.R.; Balsa-Canto, E. Structural Identifiability of Systems Biology Models: A Critical Comparison of Methods. *PLoS ONE* **2011**, *6*, e27755. [CrossRef] [PubMed]
5. Gutenkunst, R.N.; Waterfall, J.J.; Casey, F.P.; Brown, K.S.; Myers, C.R.; Sethna, J.P. Universally Sloppy Parameter Sensitivities in Systems Biology Models. *PLoS Comput. Biol.* **2007**, *3*, e189. [CrossRef] [PubMed]
6. Raue, A.; Kreutz, C.; Maiwald, T.; Bachmann, J.; Schilling, M.; Klingmüller, U.; Timmer, J. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics* **2009**, *25*, 1923–1929. [CrossRef] [PubMed]

7. Taniguchi, Y.; Choi, P.J.; Li, G.W.; Chen, H.; Babu, M.; Hearn, J.; Emili, A.; Xie, X.S. Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **2010**, *329*, 533–538. [CrossRef] [PubMed]
8. Munsky, B.; Trinh, B.; Khammash, M. Listening to the noise: Random fluctuations reveal gene network parameters. *Mol. Syst. Biol.* **2009**, *5*, 318. [CrossRef] [PubMed]
9. Zechner, C.; Ruess, J.; Krenn, P.; Pelet, S.; Peter, M.; Lygeros, J.; Koeppl, H. Moment-based inference predicts bimodality in transient gene expression. *PNAS* **2012**, *109*, 8340–8345. [CrossRef] [PubMed]
10. Helmke, U.; Hüper, K.; Khammash, M. Global identifiability of a simple linear model for gene expression analysis. In Proceedings of the 52nd IEEE CDC, Florence, Italy, 10–13 December 2013.
11. Cho, K.H.; Choo, S.M.; Jung, S.; Kim, J.R.; Choi, H.S.; Kim, J. Reverse engineering of gene regulatory networks. *IET Syst. Biol.* **2007**, *1*, 149–163. [CrossRef] [PubMed]
12. Markowetz, F.; Spang, R. Inferring cellular networks: A review. *BMC Bioinform.* **2007**, *28*, S5. [CrossRef] [PubMed]
13. Hasenauer, J.; Waldherr, S.; Doszczak, M.; Radde, N.; Scheurich, P.; Allgower, F. Identification of models of heterogeneous cell populations from population snapshot data. *BMC Bioinform.* **2011**, *12*, 125. [CrossRef] [PubMed]
14. Paulsson, J. Models of stochastic gene expression. *Phys. Life Rev.* **2005**, *2*, 157–175. [CrossRef]
15. Thattai, M.; van Oudenaarden, A. Intrinsic noise in gene regulatory networks. *PNAS* **2001**, *98*, 8614–8619. [CrossRef] [PubMed]
16. Hespanha, J. Modelling and analysis of stochastic hybrid systems. *IEE Proc. Control Theory Appl.* **2006**, *153*, 520–535. [CrossRef]
17. Sotiropoulos, V.; Kaznessis, Y. Analytical Derivation of Moment Equations in Stochastic Chemical Kinetics. *Chem. Eng. Sci.* **2011**, *66*, 268–277. [CrossRef] [PubMed]
18. Cinquemani, E. Reconstruction of promoter activity statistics from reporter protein population snapshot data. In Proceedings of the 54th IEEE CDC, Osaka, Japan, 15–18 December 2015; pp. 1471–1476.
19. Cinquemani, E. Structural identification of biochemical reaction networks from population snapshot data. In Proceedings of the 20th IFAC World Congress, IFAC—PapersOnLine, Toulouse, France, 9–14 July 2017; Volume 50, pp. 12629–12634.
20. Berthoumieux, S.; Brilli, M.; Kahn, D.; de Jong, H.; Cinquemani, E. On the identifiability of metabolic network models. *J. Math. Biol.* **2013**, *67*, 1795–1832. [CrossRef] [PubMed]
21. Bansal, M.; Belcastro, V.; Ambesi-Impiombato, A.; di Bernardo, D. How to infer gene networks from expression profiles. *Mol. Syst. Biol.* **2007**, *3*, 78. [CrossRef] [PubMed]
22. Gardner, T.; Faith, J. Reverse-engineering transcription control networks. *Phys. Life Rev.* **2005**, *2*, 65–88. [CrossRef] [PubMed]
23. Porreca, R.; Cinquemani, E.; Lygeros, J.; Ferrari-Trecate, G. Identification of genetic network dynamics with unate structure. *Bioinformatics* **2010**, *26*, 1239–1245. [CrossRef] [PubMed]
24. Neuert, G.; Munsky, B.; Tan, R.; Teytelman, L.; Khammash, M.; van Oudenaarden, A. Systematic Identification of Signal-Activated Stochastic Gene Regulation. *Science* **2013**, *339*, 584–587. [CrossRef] [PubMed]
25. Gillespie, D. A Rigorous Derivation of the Chemical Master Equation. *Physica A* **1992**, *188*, 404–425. [CrossRef]
26. Van Kampen, N. *Stochastic Processes in Physics and Chemistry*; North-Holland Personal Library: Amsterdam, The Netherlands, 1992.
27. Gadgil, C.; Lee, C.; Othmer, H. A stochastic analysis of first-order reaction networks. *Bull. Math. Biol.* **2005**, *67*, 901–946. [CrossRef] [PubMed]
28. Gillespie, D.T. The chemical Langevin equation. *J. Chem. Phys.* **2000**, *113*, 297–306. [CrossRef]
29. Gillespie, C. Moment-closure approximations for mass-action models. *IET Syst. Biol.* **2009**, *3*, 52–58. [CrossRef] [PubMed]
30. Parise, F.; Ruess, J.; Lygeros, J. Grey-box techniques for the identification of a controlled gene expression model. In Proceedings of the ECC, Strasbourg, France, 24–27 June 2014.
31. Walter, E.; Pronzato, L. *Identification of Parametric Models—From Experimental Data*; Springer: London, UK, 1997.
32. Walter, E. (Ed.) *Identifiability of Parametric Models*; Pergamon Press: Oxford, UK, 1987.

33. Khalil, H.K. *Nonlinear Systems*; Prentice Hall: Upper Saddle River, NJ, USA, 2002.
34. Ruess, J.; Lygeros, J. Identifying stochastic biochemical networks from single-cell population experiments: A comparison of approaches based on the Fisher information. In Proceedings of the 52nd IEEE CDC, Florence, Italy, 10–13 December 2013; pp. 2703–2708.
35. Kay, S.M. *Fundamentals of Statistical Signal Processing [Volume I] Estimation Theory*; Prentice Hall: Upper Saddle River, NJ, USA, 1993; p. 1.
36. De Jong, H.; Ranquet, C.; Ropers, D.; Pinel, C.; Geiselmann, J. Experimental and computational validation of models of fluorescent and luminescent reporter genes in bacteria. *BMC Syst. Biol.* **2010**, *4*, 55. [CrossRef] [PubMed]
37. Kaern, M.; Elston, T.C.; Blake, W.J.; Collins, J.J. Stochasticity in gene expression: From theories to phenotypes. *Nat. Rev. Gen.* **2005**, *6*, 451–464. [CrossRef] [PubMed]
38. Sanft, K.R.; Wu, S.; Roh, M.; Fu, J.; Lim, R.K.; Petzold, L.R. StochKit2: Software for discrete stochastic simulation of biochemical systems with events. *Bioinformatics* **2011**, *27*, 2457–2458. [CrossRef] [PubMed]
39. Ljung, L. *System Identification: Theory for the User*; Prentice Hall: Upper Saddle River, NJ, USA, 1999.
40. Callier, F.; Desoer, C. *Linear System Theory*; Springer: New York, NY, USA, 1991.
41. Boyd, S.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press: New York, NY, USA, 2004.
42. Singh, A.; Hespanha, J. Approximate Moment Dynamics for Chemically Reacting Systems. *IEEE Trans. Autom. Control* **2011**, *56*, 414–418. [CrossRef]
43. Ruess, J.; Milias-Argeitis, A.; Summers, S.; Lygeros, J. Moment estimation for chemically reacting systems by extended Kalman filtering. *J. Chem. Phys.* **2011**, *135*, 165102. [CrossRef] [PubMed]