Structural identification of biochemical reaction networks from population snapshot data

Eugenio Cinquemani

Structural identification of biochemical reaction networks from population snapshot data

Eugenio Cinquemani *

* *Inria Grenoble – Rhône-Alpes, Montbonnot, France (e-mail: eugenio.cinquemani@inria.fr).*

Abstract: In this paper we investigate how randomness in biochemical network dynamics improves identification of the network structure. Focusing on the case of so-called population snapshot data, we set out the problem as that of reconstructing the unknown stoichiometry matrix and rate parameters of the network in the case of state-affine reaction rates. We discuss what additional information is conveyed by the observation of second-order moments of the system species relative to the sole knowledge of their mean profiles. We then illustrate the impact of this additional piece of information in the reconstruction of an unknown network structure by means of a simple numerical example.

Keywords: Intrinsic noise, Genetic regulatory network, Stochastic process, Identifiability

1. INTRODUCTION

At the level of intracellular dynamics, biochemical network noise is the subject of intense research (Bowsher et al., 2013; Thattai and van Oudenaarden, 2001; Zechner et al., 2014; Llamosi et al., 2014). Intrinsic noise in gene expression is one important source of variability over different cells with identical genome, and is at the heart of the ability of microorganisms to survive changing environments, differentiate, and on (Rao et al., 2002; Kaern et al., 2005). With the advent of single-cell monitoring techniques such as flow-cytometry and videomicroscopy and their coupling with microfluidic devices and barcoding (Taniguchi et al., 2010; Klein et al., 2016; Zechner et al., 2012), randomness of single-cell responses can be observed and statistically quantified.

It is known that variability due to intrinsic noise can improve estimation of unknown parameters of biochemical reaction networks. In particular, observed second-order statistics of gene expression over multiple individuals may render identifiable parameters that would otherwise not be distinguishable from mean gene expression profiles (Munsky et al., 2009; Cinquemani, 2015). To the best of our knowledge, however, this concept has not been used systematically to address reconstruction of networks with unknown structure, except for a few efforts in model discrimination and selection (Neuert et al., 2013; Ocone et al., 2015; Lillacci and Khammash, 2010; Cinquemani et al., 2009).

In this paper we wish to investigate this point, *i.e.*, how statistics of gene expression variability can ameliorate reconstruction of unknown network structures over methods using deterministic models and based purely on mean expression profiles (as *e.g.* in Bansal et al. (2007); Porreca et al. (2010)). To this aim we consider a standard Continuous-Time Markov Chain (CTMC) model of stochastic biochemical dynamics (Paulsson, 2005; Samad et al., 2005), and focus on the dynamics of mean and variance as described by the so-called moment equations (Hespanha, 2006). We then express network reconstruction as the problem of simultaneously estimating stoichiometry and rates of the network reactions. Under the assumption of state-affine reaction rates, we discuss what information is contributed by second-order moments (*i.e.* by random noise) that is not found in the mean dynamics. We then show by way of a numerical example how indeed this additional information may circumscribe the set of network structures compatible with the data down to a much smaller pool of alternatives. To do this we rely on a two-step identification procedure similar in spirit to Parise et al. (2014), where the problem is however limited to parameter estimation in presence of a known stoichiometry.

The paper is a first investigation of the problem of stochastic network reconstruction from so-called population snapshot data (statistics of the process variables measured from independent samples taken at different time points of a dynamical experiment (Hasenauer et al., 2011)). Randomness is assumed to be primarily due to random occurrence of reactions (intrinsic noise), whereas other sources of noise such as parameter variability (extrinsic noise) are admittedly not accounted for. While arising from the network reconstruction problem in the context of gene expression, the methods and concepts discussed here are applicable to any stochastic biochemical network observed by snapshot statistics. For the fundamental case of state-affine rates, this work provides the bases for a full-blown analysis of network identifiability, and for the development of methods applicable to real data.

Stochastic modelling of biochemical reaction networks is concisely reviewed in Section 2. On this ground, the identification problem of our concern is discussed in Section 3 as follows. In Section 3.1, moment equations are rewritten in

a state-space form that is conveniently expressed in terms of stoichiometry matrix and rate parameter factors. This is used to split network structure identification into two steps. The first step, discussed in Section 3.2, concerns the reconstruction of the transition matrix of the moment dynamics from relevant observations. The specific contributions of stoichiometry and rate parameters are isolated out in a second step, as discussed in Section 3.3, where the additional information conveyed by the observation of the variance profile becomes apparent. The practical implications of this are demonstrated on a numerical example in Section 4. Conclusions and perspectives of the work are discussed in Section 5.

2. MATHEMATICAL BACKGROUND

Consider a network with n reactants and m reactions, with n known and m possibly unknown. Let $X(t) \in \mathbb{N}^n$ count the number of copies of every reactant species at a given time $t \geq 0$. For $k = 1, \ldots, m$, let $S_k \in \mathbb{Z}^n$ and $a_k(x)$, with $x \in \mathbb{N}^n$, be the stoichiometry vector and the propensity of the reaction k. That is, $\mathbb{P}[X(t + dt) = X(t) + S_k | X(t) = x] = a_k(x)dt + o(dt)$. Assuming that the probability of multiple reactions occurring in the infinitesimal time is negligible (*i.e.* $o(dt^2)$), process $X(t)$ is Markov and the evolution of the probability $\mathbb{P}[X(t) = x]$ is described by the Chemical Master Equation (Gillespie, 1992).

Now assume that propensities are affine in x, that is,

$$a_k(x) = \sum_{j=1}^{n} w_{k,j} x_j + w_{k,0}, \quad k = 1, \ldots, m.$$

This is the case for instance in networks comprising first-order reactions only. We will typically consider $w_{k,j} \geq 0$ for all indices. Let $S = [S_1 \cdots S_m] \in \mathbb{Z}^{n \times m}$ be the stoichiometry matrix of the system and define $W \in \mathbb{R}^{m \times n}$ and $w_0 \in \mathbb{R}^m$ by $(W)_{k,j} = w_{k,j}$ and $(w_0)_k = w_{k,0}$, with $k = 1, \ldots, m$ and $j = 1, \ldots, n$. Denote $\mu(t) = \mathbb{E}[X(t)]$ and $\Sigma(t) = \mathrm{Var}(X(t)) = \mathbb{E}[(X(t) - \mu(t))(X(t) - \mu(t))^T]$. It can be shown that the evolution of μ and Σ over time obeys the so-called moment equations (Hespanha, 2006),

$$\dot{\mu}(t) = SW\mu(t) + Sw_0, \tag{1}$$

$$\dot{\Sigma}(t) = SW\Sigma(t) + \Sigma(t)W^T S^T + S\mathrm{diag}(W\mu + w_0)S^T, \tag{2}$$

with initial conditions $\mu(0) = \mu_0$ and $\Sigma(0) = \Sigma_0$ determined by the initial probability distribution of X. (If propensities are not affine, then $\dot{\mu}$ and $\dot{\Sigma}$ depend on higher-order moments, and a similar system of ODEs can be obtained via moment closure techniques, see *e.g.* Zechner et al. (2012) and references therein). These equations may also accommodate the presence of a (known) forcing input driving some of the reaction rates in a time-varying fashion. This is simply obtained by replacing the relevant rate constants $w_{k,j}$ with their time-varying versions $w_{k,j}(t)$.

3. NETWORK IDENTIFICATION

The problem we consider is the reconstruction of S and W from time-course, possibly partial measurements of μ and Σ. The solution we propose relies on a convenient state-space arrangement of the moment equations and proceeds in two steps. In the first step, the state-space moment dynamics are reconstructed from the data. In the second step, a factorization problem is addressed where the specific contributions of S and W are isolated out. To achieve this we will first derive explicit expressions for the state-space system matrices in terms of S and W (Section 3.1) and outline the solution of the first step in terms of standard identification techniques (Section 3.2). We will then elaborate on the second step (Section 3.3), where the answer to the central question of this work, *i.e.* how observing randomness helps identifying S and W, is rooted.

3.1 Moment dynamics in vectorized form

Expressions (1) and (2) make up a system of linear differential equations, with the elements of μ and Σ as state variables. It is convenient to rearrange these equations in a vectorized form. Let $\underline{\Sigma} = \mathrm{vec}(\Sigma) \in \mathbb{R}^{n^2}$ and $\dot{\underline{\Sigma}} = \mathrm{vec}(\dot{\Sigma}) \in \mathbb{R}^{n^2}$. Then, using the properties of the Kronecker product (indicated by $\otimes$) to express $SW\Sigma(t)$ and its transpose in vector form, one gets that

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

where dependencies on time t are omitted from notation for brevity. For the rightmost term of the expression above, note that

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

where W_j is the j-th column of W. In view of this, using the fact that vectorization is a linear transformation, Eq. (1) and (2) may be rewritten as

$$\frac{d}{dt}\begin{bmatrix}\mu \\ \underline{\Sigma}\end{bmatrix} = \begin{bmatrix}A_{1,1} & 0_{n \times n^2} \\ A_{2,1} & A_{2,2}\end{bmatrix} \cdot \begin{bmatrix}\mu \\ \underline{\Sigma}\end{bmatrix} + \begin{bmatrix}Sw_0 \\ \mathrm{vec}(S\mathrm{diag}(w_0)S^T)\end{bmatrix}$$

with $A_{1,1} = SW$,

$$A_{2,1} = \left[\mathrm{vec}(S\mathrm{diag}(W_1)S^T) \cdots \mathrm{vec}(S\mathrm{diag}(W_n)S^T)\right],$$

$$A_{2,2} = I_n \otimes (SW) + (SW) \otimes I_n,$$

where I_n denotes the size-n identity matrix and $0_{n \times n^2}$ denotes a matrix of zeros of size $n \times n^2$. Finally, define

$$S^{(2)} = \left[\mathrm{vec}(S_1 S_1^T) \cdots \mathrm{vec}(S_m S_m^T)\right].$$

Observing that, for any m-dimensional vector w,

$$\mathrm{vec}(S\mathrm{diag}(w)S^T) = S^{(2)}w$$

we get to the following result.

Proposition 1. It holds that

$$\frac{d}{dt}\begin{bmatrix}\mu \\ \underline{\Sigma}\end{bmatrix} = \begin{bmatrix}SW & 0_{n \times n^2} \\ S^{(2)}W & A_{2,2}\end{bmatrix} \cdot \begin{bmatrix}\mu \\ \underline{\Sigma}\end{bmatrix} + \begin{bmatrix}Sw_0 \\ S^{(2)}w_0\end{bmatrix}. \tag{3}$$

Of course, due to the symmetry of covariance matrices, a smaller system of equations could be considered by focusing *e.g.* on the upper-triangular entries of Σ, at the price of some technical complicacy. In the interest of clarity, we will not detail the corresponding equations but comment on this point when needed.

3.2 Reconstruction of the moment dynamics

Let y denote the vector of p observed moments among μ and Σ. Calling ξ the state vector of (3), the state-space model of the observed system is then

$$\dot{\xi}(t) = A\xi(t) + \omega, \qquad t \geq 0, \qquad \xi(0) = \xi_0, \quad (4)$$
$$y(t) = C\xi(t), \quad (5)$$

where

$$A = \begin{bmatrix} SW & 0_{n\times n^2} \\ S^{(2)}W & A_{2,2} \end{bmatrix}, \qquad \omega = \begin{bmatrix} Sw_0 \\ S^{(2)}w_0 \end{bmatrix}. \quad (6)$$

The specific form of observation matrix $C \in \mathbb{R}^{p\times(n+n^2)}$ determines what statistics are actually observed, and is fixed by the experimental setup. Additionally, an exogenous input u is typically present. For typical biological scenarios, this could be a known input driving one of the rates composing w_0 (Parise et al., 2014). In this case term ω may be rewritten as $Bu(t)$, with B unknown and some entries of u set to 1 to account for constant nonzero entries of w_0. In practice, the system outputs $y(t)$ are typically observed at sample times in some finite set $\mathscr{T}$, yielding noisy mesurements $\tilde{y}(t) = y(t) + e(t)$, with $t \in \mathscr{T}$, where $e(t)$ is the measurement error.

Identification of such dynamical system can be addressed by standard identification techniques (Ljung, 1999; Walter and Pronzato, 1997) and may take advantage of structural properties of the system (4), such as the block-triangular structure of matrix A. We will make the following assumption.

Assumption 1. The minimal realization of the input-output map taking u into y is of order at least p. In addition, the class of minimal realizations is uniquely identifiable given u and $y(t)$, $t \in \mathscr{T}$.

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. On this basis, the second part of the assumption concerns the informativity of the experiment, *i.e.* it requires that the available data is sufficiently rich to make identification of the moment dynamics well-posed. The role of noise (which does not enter Assumption 1), is not the main focus of our discourse and we will come back to it later. Instead note that, if (A^*, B^*, C^*) is one minimal realization, all minimal realizations are of the form $(TA^*T^{-1}, TB^*, C^*T^{-1})$ for a nonsingular square matrix T. That is, under Assumption 1, identification is well-posed up to an unknown factor T. However, T is constrained by the knowledge of C.

In particular we consider two cases:

Case (i): Observation of μ only. This is the case for $C = \begin{bmatrix} C' & 0_{p'\times n^2} \end{bmatrix}$ with $C' = \mathrm{diag}(c_1, \ldots, c_{p'})$, where $p' = n$, and $c_\ell > 0$, with $\ell = 1, \ldots, p'$. Due to the block-triangular structure of A, in this case, the observed dynamics can be rewritten as

$$\dot{\xi}'(t) = SW\xi'(t) + B'u,$$
$$y(t) = C'\mu(t).$$

This is a realization of the system of order p'. Under Assumption 1, this is also a minimal realization that is well-determined by the data up to an unknown factor T. For any minimal realization (A^*, B^*, C^*), it must then hold that $C' = C^*T^{-1}$, *i.e.* $T = C^*C'^{-1}$. Then $SW = TA^*T^{-1}$ is also uniquely determined.

Case (ii): Observation of μ and Σ. This case is captured by a selection matrix C with $p'' = n + n(n+1)/2$ rows and $n+n^2$ columns. Every column has at most one nonzero entry, and every row has exactly one nonzero entry in a position corresponding to either one element of μ, or one element of the upper-triangular part of Σ. Then, recalling that $\Sigma = \Sigma^T$, system (4)–(5) may be rewritten as

$$\dot{\xi}''(t) = A''\xi''(t) + B''u,$$
$$y(t) = C''\xi''(t)$$

with $A \in \mathbb{R}^{p''\times p''}$, $C'' = \mathrm{diag}(c_1, \ldots, c_{p''})$ and $c_\ell > 0$, with $\ell = 1, \ldots, p''$. This is a realization of the system of order p'' and, by virtue of Assumption 1 and the same arguments as for case (i), matrix A'' is uniquely determined. Clearly A'' contains all elements of the full matrix A in (6). Therefore, also A is uniquely determined.

To sum up, under Assumption 1, Case (i) yields unique reconstruction of $A_{1,1} = SW$, while Case (ii) also yields unique reconstruction of $A_{2,1}$ and $A_{2,2}$. For a specific input configuration $\omega = Bu$, the reconstruction of matrix B could also be discussed. However, for our subsequent analysis, the information that this may carry about S will not be taken into consideration.

Other cases of interest of course exist, *i.e.* only the statistics of a subset of the state variables X are observed, or only the diagonal of Σ is observed (Cinquemani, 2015; Munsky et al., 2009; Zechner et al., 2012). The role of Assumption 1 in the uniqueness of the identified dynamics becomes then more involved and requires investigation out of the scope of this paper.

3.3 Reconstruction of network structure and parameters

We now assume that the system dynamics A have been determined as in the previous section. The network reconstruction task becomes that of extracting S and W from A. Different information about A is available. For Case (i) , *i.e.* observation of the mean, only the product SW is known. For Case (ii), *i.e.* observation of mean and covariance matrix, the full matrix A is known. The original question, that is, what can be said about S and W upon observation of higher-order statistics that cannot be said by observation of the sole process mean, becomes that of understanding in what way the full knowledge of A helps the factorization of SW into S and W relative to the sole knowledge of the product SW.

We start by assuming that the number of network reactions m is known. By inspecting the structure of A, it can be appreciated that block $A_{2,2}$ does not add any information, since S and W enter $A_{2,2}$ only by their product SW. What provides additional information is block $A_{2,1}$, responsible for the coupling between the dynamics of μ and that of Σ. Indeed the elements of S and W appear in $A_{2,1}$ via the product $S^{(2)}W$, and the elements of $S^{(2)}$ are not the same as but rather cross-products of the elements of S. Suppose thus that estimates $\hat{A}_{1,1}$ of SW and $\hat{A}_{2,1}$ of $A_{2,1}$ are available. Characterizing the solutions of the factorization of $\hat{A}_{1,1}$ into S and W that are compatible with $\hat{A}_{2,1} = S^{(2)}W$ appears highly nontrivial, partly due to the discrete nature of the entries of S. We can, however, describe the solution in terms of an optimization problem.

Let $\underline{W} = \text{vec}(W)$ and

$$\hat{z} = \left[\text{vec}(\hat{A}_{1,1})^T \ \text{vec}(\hat{A}_{2,1})^T\right]^T, \quad (7)$$

$$z = \left[\text{vec}(A_{1,1})^T \ \text{vec}(A_{2,1})^T\right]^T. \quad (8)$$

In this context, $\hat{z}$ represents data, while z is a function of S and W. In particular, it is a linear function of W that can be written as

$$z = Z(S)\underline{W}, \quad Z(S) = \begin{bmatrix} I_n \otimes S \\ I_n \otimes S^{(2)} \end{bmatrix}.$$

Now, for the identification of S and $\underline{W}$, consider the optimization problem

$$\min_{S \in \mathbb{Z}^{n\times m}, W \in \mathbb{R}^{m\times n}_{\geq 0}} \left\| M\left(\hat{z} - Z(S)\underline{W}\right)\right\| \quad (9)$$

where $\|\cdot\|$ denotes Euclidean norm and M is a selection matrix. For Case (ii) (simultaneous observations of μ and Σ), taking $M = I_{n^2+n^3}$, the solution of this problem is any couple (S, W) matching the estimated blocks $\hat{A}_{1,1}$ and $\hat{A}_{2,1}$. For Case (i) (observation of μ only), taking $M = [I_{n^2} \ 0_{n^2\times n^3}]$, the solution is any couple (S, W) matching $\hat{A}_{1,1}$ only.

For any choice of M, problem (9) is a Mixed-Integer Non-Linear Program (MINLP), which is per se hard to solve. We can however rewrite the problem as

$$\min_{S\in\mathbb{Z}^{n\times m}} Q(S), \quad Q(S) = \min_{W\in\mathbb{R}^{m\times n}_{\geq 0}} \|M\left(\hat{z} - Z(S)\underline{W}\right)\|.$$

The innermost optimization problem is a Linearly constrained Quadratic Program (LQP), whence it is convex and can be solved very efficiently (Boyd and Vandenberghe, 2004). For any given S, let $\hat{Q}(S)$ be the minimum value of the LQP and $\underline{\hat{W}}(S)$ be a solution, *i.e.* $\hat{Q}(S) = \left\|M\left(\hat{z} - Z(S)\underline{\hat{W}}(S)\right)\right\|$. Taking $\hat{Q}(S)$ as the cost for a given S, overall solutions to (9) can now be sought by solving

$$\hat{S} \in \arg\min_{S\in\mathbb{Z}^{n\times m}} \hat{Q}(S). \quad (10)$$

This is an integer optimization problem. For, say, $|S_{i,k}| \leq S_{max}$, with $i = 1, \ldots, n$ and $k = 1, \ldots, m$, it can be solved by enumeration. Despite the exponential complexity of the enumeration ($O(2S_{max}+1)^{n\cdot m}$), this approach is exact and still convenient for networks of moderate size. For a solution $\hat{S}$, the corresponding estimate of W is then $\underline{\hat{W}}(\hat{S})$.

For perfect estimates $\hat{A}_{1,1}$ and $\hat{A}_{2,1}$, subsequent estimates $\hat{S}$ and $\hat{W}$ would attain zero residual error $\hat{Q}(\hat{S})$ at least for the true network structure and parameters (S, W). Solutions attaining zero error may not be unique, *i.e.* a pool of undistinguishable solutions is generally obtained. Yet, the additional observation of Σ, *i.e.* the availability of $\hat{A}_{2,1}$, is expected to yield a much smaller pool of indistinguishable solutions compared to what can be obtained by the observation of the sole mean, since Problem (9) is more constrained (M has more rows). These considerations will be demonstrated by an example in the next section.

In practice, acknowledging the presence of unavoidable measurement noise e, estimates $\hat{A}_{1,1}$ and $\hat{A}_{2,1}$ will themselves be affected by estimation error. In that case, estimates of S and W attaining zero cost Q are not to be expected. The set of practically indistinguishable stoichiometry matrices should then rather be defined as

$$\{\hat{S} : \ \hat{Q}(\hat{S}) < \epsilon\} \quad (11)$$

along with corresponding parameter estimates $\underline{\hat{W}}(\hat{S})$, for a small ϵ to be set as a function of the magnitude of the estimation error affecting $\hat{A}_{1,1}$ and $\hat{A}_{2,1}$.

Finally, consider the case where the true number of reaction channels m is unknown. One way to address the problem is to carry out the identification above for a sufficiently large value of m. As a result of this procedure, let $\hat{m} \leq m$ be the minimal number of nonzero columns in the pool of stoichiometry matrices $\hat{S}$ compatible with the data. Then, this value clearly represents the minimal number of reaction channels needed to explain the data, since the columns of $\hat{S}$ that are identically zero do not contribute to the network dynamics and can be eliminated from the solution. In practice, this is a wasteful procedure, since it implies the exploration of an unnecessarily large set of options for S. A better procedure is to incrementally explore matrices S of increasing column size m, and to stop the search when a nonempty pool of solutions explaining the data is found. We will again exemplify the results of this procedure in the next section.

4. NUMERICAL CASE STUDY

We now investigate the results of Section 3.3 by means of a simple numerical example, verifying in particular to what extent observations of the covariance matrix Σ reduce the number of indistinguishable solutions S and W. Consider a reaction network with

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

In accordance with the discussion of Section 3.2, we assume that the corresponding blocks of matrix A have been estimated without error, *i.e.* $\hat{A}_{1,1} = A_{1,1}$ for Case (i) and also $\hat{A}_{2,1} = A_{2,1}$ for Case (ii). We implemented in Matlab the estimation of S and W by solving (10) via enumeration, for $S_{max} = 2$. For every candidate stoichiometry S, the inner optimization problem was solved via the Matlab function `lsqnonneg`, implementing quadratic optimization under nonnegativity constraints. To cope with numerical errors in the solution of the inner optimization, indistinguishable solutions compatible with the data have been selected with the approach of Eq. (11) with an appropriately small $\epsilon = 10^{-6}$.

We first tested the approach by fixing the number of reaction channels to the true value $m = 3$. There are thus a total of $(2S_{max}+1)^{n\cdot m} = 5^6 = 15625$ possible stoichiometry matrices. The exploration of all such alternatives takes about five seconds on an Intel i7 processor. Results are summarized in Table 1.

In Case (i) (observation of mean only), matching the observed block $\hat{A}_{1,1}$ exactly (within ϵ error) returned 2604 solutions $(\hat{S}, \hat{W})$, *i.e.* about 16.7% of all stoichiometries tested can explain the data, provided an adapted choice of rate parameters. In Case (ii) (observation of mean and covariance), the requirement to simultaneously match blocks $\hat{A}_{1,1}$ and $\hat{A}_{2,1}$ lead to only 6 possible solutions, *i.e.* only about 0.038% of all possible stoichiometries is in agreement with the data. The gain in using observations

	m	1	2	3	4
Case (i)	#	0	4	2604	150172
	%	0.0	0.64	16.7	38.4
	T	0.022	0.22	5.5	148
Case (ii)	#	0	0	6	564
	%	0	0	0.038	0.14
	T	0.024	0.25	6.5	168

Table 1. Number of solutions (#), percent of stoichiometry matrices accepted over all matrices tested (%), and computational time in seconds (T) in the reconstruction of the example network for Case (i) (observation of μ only) and Case (ii) (observation of μ and Σ), for different hypotheses on the number of reactions (m).

of the second-order moments is thus striking. It should however be noticed that many of the alternative network structures explored, as well as many of the solutions found, are equivalent, in the sense that they are composed of permutations of the reaction list, *i.e.* permutations of the columns of S and corresponding permutation of the rows of W. At a closer look, it turns out that the solutions found in Case (ii) are all equivalent to the correct solution, in the sense that estimates $\hat{S}$ are column permutations of the true S, and are accompanied by estimates $\hat{W}$ that are row permutations of W. That is, the same 3 reactions are listed in the 6 possible different orders. Thus, in this case, the solution returned is correct and essentially unique. These 6 solutions are, on the other hand, just a small subset of the large pool of 2604 putative solutions found in Case (i).

In order to test estimation of the number of reaction channels m, we also run the algorithm for smaller hypothetical values of m, *i.e.* $m = 1$ and $m = 2$, and for a larger hypothetical value of $m = 4$. For Case (i), solutions are found starting from $m = 2$. This is consistent with the fact that the columns of SW are vectors in a two-dimensional space, so that at least two columns of $\hat{S}$ are needed to span this space, *i.e.* to match the columns of SW via the product $\hat{S}\hat{W}$ by an appropriate choice of $\hat{W}$. However, this shows that Case (i) does not allow for the estimation of the true m. On the other hand, in Case (ii), solutions are found starting from the true value $m = 3$, whereas simpler solutions comprising fewer reactions are ruled out. When instead run for a maximal hypothetical value of $m = 4$, out of all $5^8 = 390625$ stoichiometries tested, many solutions are found in both cases, including the correct solution in the form of stoichiometries $\hat{S}$ with one column of zeroes and the remaining columns equal to a permutation of the columns of the true S. Many alternative and incorrect solutions are also found due to the over-parametrization of mean and variance dynamics associated with the overly large value of m. Still, Case (ii) remains much more selective (compare entries # and % in Table 1).

The computational time was in the same order for Case (i) and Case (ii) and increases rapidly with m, as expected. In particular, the test with $m = 4$ witnesses the utility of estimating model order (number of reactions) with an incremental approach as discussed at the end of the previous section. In particular in Case (ii), exploring alternatives with increasing hypothetical value of m would incur a computational time of less than 6 seconds, since at $m = 3$ the iteration would stop with a nonempty pool of viable network structures. Exploring all solutions with $m = 4$ instead took almost three minutes, and additional postprocessing time would be required to figure out the correct (smaller) model order $m = 3$ from the pool of viable solutions.

5. CONCLUSIONS

In this work we have discussed identification of the structure of biochemical networks in the sense of estimating both stoichiometry and rates of the unknown network reactions. Focusing on the case of state-affine reaction rates, we have shown that observation of second-order moments of the network species allows one to dramatically reduce the pool of solutions that would be found from the observation of mean profiles only. To achieve this, we have proposed a two-step procedure, where a preliminary step is devoted to reconstruct the moment dynamics from which stoichiometry and rates are isolated out in a second step. We have also discussed an incremental approach to estimate the number of reactions composing the network, and verified methods and results by way of a simple numerical example.

The results obtained here provide the basis for the full development of theory and methods for structural identification of unknown reaction networks from population snapshot data. A number of directions of investigation are open. Concerning the identification procedure, in particular, the following questions need to be addressed: For the first step (Section 3.2), the investigation of other measurement models of interest, *e.g.* the identifiability of the second-order moment dynamics in absence of readouts for the covariance among different network species; For the second step (Section 3.3), the algebraic characterization of the space of solutions for a given state matrix of the moment dynamics; And for both steps, the analysis of practical identifiability, *i.e.* distinguishability and reconstruction accuracy of the moment dynamics as well as of reaction stoichiometry and rates from noisy and finite sample datasets. In a broader perspective, a relevant challenge is the exploration of possible conceptual and methodological generalizations of the results to non-affine reaction rate functions. Applications to real biological data and systems is of course the ultimate goal of the research.

REFERENCES

Bansal, M., Belcastro, V., Ambesi-Impiombato, A., and di Bernardo, D. (2007). How to infer gene networks from expression profiles. *Mol. Syst. Biol.*, 3(78).

Bowsher, C.G., Voliotis, M., and Swain, P.S. (2013). The fidelity of dynamic signaling by noisy biomolecular networks. *PLoS Comput Biol*, 9(3), e1002965.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Cinquemani, E. (2015). Reconstruction of promoter activity statistics from reporter protein population snapshot data. In *2015 54th IEEE Conference on Decision and Control (CDC)*, 1471–1476.

Cinquemani, E., Milias-Argeitis, A., Summers, S., and Lygeros, J. (2009). Local identification of piecewise deterministic models of genetic networks. In R. Majumdar and P. Tabuada (eds.), *HSCC*, volume 5469 of *LNCS*, 105–119. Springer.

Gillespie, D. (1992). A rigorous derivation of the chemical master equation. *Physica A: Statistical Mechanics and its Applications*, 188(1), 404 – 425.

Hasenauer, J., Waldherr, S., Doszczak, M., Radde, N., Scheurich, P., and Allgower, F. (2011). Identification of models of heterogeneous cell populations from population snapshot data. *BMC Bioinformatics*, 12(1), 125.

Hespanha, J. (2006). Modelling and analysis of stochastic hybrid systems. *Control Theory and Applications, IEE Proceedings*, 153(5), 520–535.

Kaern, M., Elston, T.C., Blake, W.J., and Collins, J.J. (2005). Stochasticity in gene expression: From theories to phenotypes. *Nat. Rev. Gen.*, 6, 451–464.

Klein, A., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D., and Kirschner, M. (2016). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161, 1187–1201.

Lillacci, G. and Khammash, M. (2010). Parameter estimation and model selection in computational biology. *PLoS Comput. Biol.*, 6, e1000696.

Ljung, L. (1999). *System Identification: Theory for the user*. Prentice Hall.

Llamosi, A., Gonzalez-Vargas, A.M., Versari, C., Cinquemani, E., Ferrari-Trecate, G., Hersen, P., and Batt, G. (2014). What population reveals about individual cell identity: gene expression variability in yeast. *Under submission to PNAS*.

Munsky, B., Trinh, B., and Khammash, M. (2009). Listening to the noise: Random fluctuations reveal gene network parameters. *Mol. Syst. Biol.*, 5(318).

Neuert, G., Munsky, B., Tan, R., Teytelman, L., Khammash, M., and van Oudenaarden, A. (2013). Systematic identification of signal-activated stochastic gene regulation. *Science*, 339(6119), 584–587.

Ocone, A., Haghverdi, L., Mueller, N.S., and Theis, F.J. (2015). Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data. *Bioinformatics*, 31(12), i89–i96.

Parise, F., Ruess, J., and Lygeros, J. (2014). Grey-box techniques for the identification of a controlled gene expression model. In *Proceedings of the ECC*.

Paulsson, J. (2005). Models of stochastic gene expression. *Phys. Life Rev.*, 2(2), 157 – 175.

Porreca, R., Cinquemani, E., Lygeros, J., and Ferrari-Trecate, G. (2010). Identification of genetic network dynamics with unate structure. *Bioinformatics*, 26(9), 1239–1245.

Rao, C., Wolf, D., and A.P.Arkin (2002). Control, exploitation and tolerance of intracellular noise. *Nature*, 420, 231–237.

Samad, H.E., Khammash, M., Petzold, L., and Gillespie, D. (2005). Stochastic modelling of gene regulatory networks. *Int. J. Robust Nonlin. Contr.*, 15, 691–711.

Taniguchi, Y., Choi, P.J., Li, G.W., Chen, H., Babu, M., Hearn, J., Emili, A., and Xie, X.S. (2010). Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329, 533–538.

Thattai, M. and van Oudenaarden, A. (2001). Intrinsic noise in gene regulatory networks. *PNAS*, 98(15), 8614–8619.

Walter, E. and Pronzato, L. (1997). *Identification parametric models – from experimental data*. Springer.

Zechner, C., Ruess, J., Krenn, P., Pelet, S., Peter, M., Lygeros, J., and Koeppl, H. (2012). Moment-based inference predicts bimodality in transient gene expression. *PNAS*, 21(109), 8340–8345.

Zechner, C., Unger, M., Pelet, S., Peter, M., and Koeppl, H. (2014). Scalable inference of heterogeneous reaction kinetics from pooled single-cell recordings. *Nature Methods*, 11, 197–202.