Published in final edited form as:

J Coupled Syst Multiscale Dyn. 2013 December ; 1(4): 468–475.

Algebraic Statistical Model for Biochemical Network Dynamics Inference

Daniel F. Linder¹ and Grzegorz A. Rempala^{2,3,*}

¹Department of Biostatistics, Jiann-Ping Hsu College of Public Health, Georgia Southern University, P.O. Box 8015 Statesboro, GA 30460

²Division of Biostatistics, Ohio State University, Cunz Hall, 1841 Neil Ave. Columbus, OH 43210

³Mathematical Biosciences Institute, Ohio State University, Jennings Hall, 1735 Neil Ave. Columbus, OH 43210

Abstract

With modern molecular quantification methods, like, for instance, high throughput sequencing, biologists may perform multiple complex experiments and collect longitudinal data on RNA and DNA concentrations. Such data may be then used to infer cellular level interactions between the molecular entities of interest. One method which formalizes such inference is the stoichiometric algebraic statistical model (SASM) of [2] which allows to analyze the so-called conic (or single source) networks. Despite its intuitive appeal, up until now the SASM has been only heuristically studied on few simple examples. The current paper provides a more formal mathematical treatment of the SASM, expanding the original model to a wider class of reaction systems decomposable into multiple conic subnetworks. In particular, it is proved here that on such networks the SASM enjoys the so-called sparsistency property, that is, it asymptotically (with the number of observed network trajectories) discards the false interactions by setting their reaction rates to zero. For illustration, we apply the extended SASM to in silico data from a generic decomposable network as well as to biological data from an experimental search for a possible transcription factor for the heat shock protein 70 (Hsp70) in the zebrafish retina.

Keywords

Systems Biology; Biochemical Networks; Law of Mass Action; Algebraic Statistical Model; Parameter Inference; DNA- and RNA-based Technologies

1. INTRODUCTION

New DNA-based technologies are revolutionizing the systems biology approach to collecting and analyzing molecular data. These technologies are capable of producing estimates of mRNA levels in a biological sample either through estimating molecular counts (for instance, via RT-PCR, high throughput RNA-seq, or molecular flow cytometry, see,

*Corresponding author: grempala@cph.osu.edu, Tel.: +1 614 2476289, Fax: +1 614 2923572.

Conflicts of Interest Disclosure: None declared.

e.g., [9]) or their proxies like the fluorescence intensity of the DNA-microarray probes (see, e.g., [10]). However, in order to take advantage of these advancements and gain meaningful insights into the molecular mechanisms which give rise to such data, one needs to establish formal ways of reverse engineering reaction networks from the longitudinal samples of their trajectories. Whereas multiple such methods exist in the literature (see, e.g., [11]), their statistical properties are often difficult to establish. In order to partially address this issue we study here large sample properties of one such method, the so-called algebraic statistical model for biochemical reaction networks proposed in [2] and further expanded in [3].

Formally, a biochemical reaction network is a collection of chemical species along with their interactions described by the network stoichiometric matrix and the corresponding reaction rates. The network's longitudinal evolution is defined as the time evolution of all molecular concentrations of the network species. In principle, given the network stoichiometry and its reaction rate laws, one should be able to completely determine the set of nonzero rate reactions from the species concentration data collected from one or more reaction network trajectories on a sufficiently dense time grid. However, in practice the number of data points and species types observed is typically constrained by both the technological limitations and practical cost considerations. Moreover, in most cases of interest the observed quantities are distorted by (often sizable) experimental measurement errors. Compounding the difficulty is also the fact that the stoichiometric matrix of the network is often at least partially unknown and that the rate laws may be changing from experiment to experiment, due to changes in the external conditions beyond experimenters' control.

In many cases the confluence of some or all of the above issues can make the standard, likelihood-based inferential problems challenging, due to the general sensitivity of the likelihood functions to model mis-specifications. The stoichiometric algebraic statistical model (SASM) approach suggested in [2] provides a more robust approach to likelihood based inference for biochemical networks by combining the ideas of algebraic geometry with the empirical likelihood theory. Although originally the SASM was developed for conic networks, i.e., the reaction networks sharing a common source species, it is not difficult to see that the method extends rather easily to a large class of 'source decomposable' reaction models (see below). The SASM method discussed here is based on the following three basic assumptions: (i) that at least part of the network stoichiometry is known, (ii) that the reaction rate laws are changing according to a prescribed rule which is identifiable from the available data, and (iii) that the observed concentration estimates are, at least asymptotically, in the sense of large volume approximations (see below), unbiased. The first assumption is introduced so as to limit the number of possible structures considered, whereas the remaining two ensure that the inferential problem is (asymptotically) well-posed. Regarding (ii), note that since the network trajectories may be observed under varying experimental conditions, like differences in temperatures or catalysts (e.g., enzymes) concentrations, the reaction laws may vary across experiments. In the current paper the assumption (ii) is made tacitly below by considering the family of mass-action rate laws where the kinetic parameter follows a gamma family of distributions as in [3]. Extensions to other rate laws are relatively straightforward, but are not considered here. Under the identifiability assumptions, we show below that the SASM method is sparsity-

consistent or “sparsistent” in the sense that it asymptotically preserves the sparsity pattern of the network connectivity matrix in which the sets of chemical substrates and products (the so-called species complexes) are directly connected if and only if there exists a nonzero rate reaction between them.

The paper is organized as follows. We begin with a brief recap of the relevant mass action kinetics and algebraic models in Sections 2.1 and 2.2. In Section 2.3 the sparsistency result is given as Theorem 1. In Section 3 the illustration of the SASM method is given by discussing its application to a simple in silico model as well as to the rod photoreceptors network in the longitudinal analysis of the zebrafish retina data. In Section 4 we give a brief summary of our main points and offer some concluding remarks.

2. STOICHIOMETRIC ALGEBRAIC STATISTICAL MODEL

2.1. Jump Processes and Mass Action Law

Reaction systems are typically described by a set of d constituent chemical species reacting according to a finite set of equations describing how the system’s state changes in the course of a particular reaction. Using the standard chemical notation, the reaction

is interpreted as “a molecule of species A_1 combines with a molecule of A_2 to produce a molecule of A_3 at the rate λ ”. The reaction system is just a collection of all such reactions involving d species, treated as a continuous time Markov chain. The chain state space is given by all nonnegative integer vectors $X(t) = x \in \mathbb{Z}_{\geq 0}^d$ describing the molecular counts of the d species. The k -th reaction is modeled by a pair of vectors, v_k giving the number of molecules of each species consumed by the reaction and v'_k giving the number of molecules of each species produced by the reaction, along with the reaction rate functions $\lambda_k(x)$. The vectors x are elements on the lattice $\mathbb{Z}_{\geq 0}^d$ and the corresponding reaction rates are non-negative scalars. The matrix $v' - v$ with columns $v'_k - v_k$ is known in the literature as the stoichiometric matrix. If the k -th reaction occurs at time t , the system state updates according to

$$X(t) = X(t-) + v'_k - v_k.$$

Under the assumptions that the system is well-stirred and thermally equilibrated, the reaction rates take on the form of a classical law of mass action. Specifically, defining $|v_k| = \sum_i v_{ik}$ the k -th rate function have the form

$$\lambda_k^{(n)}(x) = n \kappa_k \frac{\prod_i v_{ik}!^{(x_i)}}{n^{|v_k|}} \quad (2)$$

which describes combinatorially the number of ways to select homogenous molecules needed for the k -th reaction to occur. Here and elsewhere in the paper $\kappa_k \geq 0$ are the reaction

constants dependent upon the chemical conditions of the network. In particular, $\kappa_k = 0$ indicates that the corresponding reaction does not occur.

The exponential waiting time distribution, along with the probability of a particular reaction (see [4]) justifies the expression for system's time evolution

$$X(t) = X(0) + \sum_k Y_k \left(\int_0^t \lambda_k^{(n)}(X(s)) ds \right) (v'_k - v_k), \quad (3)$$

where Y_k are independent unit Poisson processes. Here n denotes the size of the system, typically given by its volume times Avagadro's number. Scaling the species counts by n gives $c = n^{-1}x$, a vector of species concentrations in moles per unit volume. By considering the asymptotic rate functions in the large volume limit ($n \rightarrow \infty$), one may obtain an approximation of the mass action rates in terms of the species concentration rates

$$\lambda_k^{(n)}(x) \approx n \kappa_k \prod_i c_i^{v_{ik}} =: n \lambda_k(c).$$

This approximation (see, e.g., the appendix in [12]) allows one to recognize the large volume limit of the stochastic jump process (3) as a solution of the classical deterministic law of mass action system of equations

$$\dot{c}(t) = \sum_k \kappa_k c^{v_k}(t) (v'_k - v_k), \quad (4)$$

where $c^{v_k}(t) = \prod_i c_i^{v_{ik}}(t)$. The above ODEs are parametrized by the rate constants κ and possibly the reaction network initial conditions. Throughout the paper all the systems of chemical reactions (1) are meant to represent their limiting ODEs (4). The issue of estimating the ODEs parameters κ from trajectory data has been recently studied in [7, 12] by means of the least squares method. In what follows, we assume that the practically computable and consistent (like the least squares) estimates of the linear combinations of κ (cf. (6) below), are available based on the trajectory data (see, e.g., [3] for further discussion). The statistical problem of interest is to identify, based on the estimates of the unknown parameters in (4) and the stoichiometry $v' - v$, the true reactions of the network, that is, to identify the values of k for which the rate constants are strictly positive $\kappa_k > 0$.

2.2. Algebraic Multinomial Model

As it turns out, the problem of identifying all k for which $\kappa_k > 0$ may be viewed as an algebraic-geometric one, based on the so-called conic network decomposition [2]. A conic network is defined as one in which all the reactions share the same source species complex. The figure below gives an example of a conic network where v_0 molecules of species A_0 are substrates for all the reactions.

For the conic network, the equation (4) reduces to

$$\dot{c}(t) = c^{v_0}(t) \sum_k \kappa_k (v'_k - v_k) = c^{v_0}(t) \gamma, \quad (5)$$

where $c^{v_0}(t) = \prod_i c_i^{v_{i0}}(t)$ does not depend on k and

$$\gamma = \sum_k \kappa_k (v'_k - v_k) \quad (6)$$

is a column vector which does not depend on t . In general, the equation (4) is seen to admit the following decomposition into multiple (say s) conic networks with the sources

corresponding to $c_{(j)}^{v_0}(t) = \prod_i c_i^{v_{ikj}}(t)$, $j = 1, \dots, s$

$$\dot{c}(t) = \sum_{j=1}^s c_{(j)}^{v_0}(t) \sum_k \kappa_k (v'_k - v_k) = \sum_{j=1}^s c_{(j)}^{v_0}(t) \gamma_j. \quad (7)$$

From the algebraic form of the above ODEs it is clear that the search for $\kappa_k > 0$ requires the analysis of their various linear combinations, the γ_j vectors. Although so far we treated the κ_k 's as constants, in order to be able to incorporate varying experimental conditions and chemical parameters (see Section 1), it is convenient to relax this assumption and let κ_k 's be instead independently distributed with some non-negatively supported distribution. In what follows we assume that all reaction rates are given by a mixture of the gamma distribution $\Gamma(\alpha, b)$ and the degenerate distribution δ_0 which puts the unit mass at 0

$$\kappa_k \sim (1 - \theta_k) \delta_0 + \theta_k \Gamma(\alpha_k, b) \quad (8)$$

$k = 1, \dots, m$ and $\kappa_{k_i} \perp \kappa_{k_j}$. Here $\alpha_k > 0$ and the mixing coefficient $\theta_k \in [0, 1]$, with $\theta_k = 0$ implying the degenerate distribution concentrated at 0 and corresponding to a false reaction. Under (8) the system (7) formally becomes the ODEs system with random coefficients (see, e.g., [13]). As multiple experiments are performed, the different values of $\gamma \in \mathbb{R}^d$ are recorded by observing the trajectory (7). The relation (8) implies that only $\alpha_k > 0$ effect the γ values given by (6). Due to our assumption on the estimability of γ , in what follows we treat their values as observables and refer to them below as “data points” or “observations”. In order to simplify the discussion, but without losing generality, suppose now that we only

have a single source cone network (5) of m reactions $R = (R_1, \dots, R_m)$. Define $\mathcal{R}_d$ to be the collection of all $\binom{m}{d}$ positive cones spanned by the combinations of d reaction vectors in R and define $S = \{S_1, \dots, S_n\}$ to be the partition of R into disjoint sets in $\mathbb{R}^d$ obtained by taking all possible intersections of the non-degenerate cones in $\mathcal{R}_d$. These disjoint regions S_i are referred to as chambers and each data point γ belongs to exactly one of them.

Let $\mathcal{D}_d^m$ be the set of all possible selections σ of unordered d -tuples out of a given m -tuple, and let $\{\gamma \sim \mathcal{D}_d^m\}$ denote the event that the right-hand side of the ODEs (5) has only d non-zero rates present, to which we shall also refer below for brevity as “ γ being generated by a d -tuple of reactions”. Further, let $\{\gamma \sim \sigma(1) \dots \sigma(d)\}$ denote the event that γ was generated by the particular d -tuple, $\sigma(1) \dots \sigma(d)$ and let C_σ stand for the cone generated by the reactions $R_{\sigma(1)}, \dots, R_{\sigma(d)}$. The event $\{\gamma \sim \mathcal{D}_d^m\}$ may be written as the disjoint union

$$\{\gamma \sim \mathcal{D}_d^m\} = \bigcup_{C_\sigma \in \mathcal{R}_d} \{\{\gamma \in C_\sigma\} \cap \{\gamma \sim \sigma(1) \dots \sigma(d)\}\}, \quad (9)$$

where the set summation is taken over all possible cones $C_\sigma \in \mathbb{R}^d$ or, which is the same, over all $\sigma \in \mathcal{D}_d^m$. It is perhaps worthy to note that, due to (8), for any d -tuple $\sigma(1) \dots \sigma(d)$,

conditionally on the event of probability $\prod_{k=1}^d \theta_{\sigma(k)} > 0$, the random vector

$$(\bar{\kappa}_{\sigma(1)}, \dots, \bar{\kappa}_{\sigma(d)}) \text{ where } \bar{\kappa}_{\sigma(k)} = \frac{\kappa_{\sigma(k)}}{\sum_{i=1}^d \kappa_{\sigma(i)}}, \text{ has the Dirichlet distribution } Dir(a_{\sigma(1)}, \dots, a_{\sigma(d)}).$$

Set $\theta = (\theta_1, \dots, \theta_m)$ and consider now the probability $g_i(\theta)$ that a data point γ generated by a d -tuple falls into the chamber S_i . Then for $i = 1, \dots, m$

$$\begin{aligned} g_i(\theta) &= P((\gamma \in S_i) \cap (\gamma \sim \mathcal{D}_d^m)) \\ &= P(\bigcup_{C_\sigma \in \mathcal{R}_d} \{\{\gamma \in C_\sigma \cap S_i\} \cap \{\gamma \sim \sigma(1) \dots \sigma(d)\}\}) \\ &= \sum_{C_\sigma \in \mathcal{R}_d} P(\{\gamma \in C_\sigma \cap S_i\} | \{\gamma \sim \sigma(1) \dots \sigma(d)\}) \times P(\{\gamma \sim \sigma(1) \dots \sigma(d)\}) \quad (10) \\ &= \sum_{C_\sigma \in \mathcal{R}_d} \frac{\text{vol}(C_\sigma \cap S_i)}{\text{vol}(C_\sigma)} \theta_{\sigma(1)} \dots \theta_{\sigma(d)}. \end{aligned}$$

Here the third equality follows from the disjointness of events and Bayes' theorem, and the quantities $\text{vol}(C_\sigma)$ and $\text{vol}(C_\sigma \cap S_i)$ are defined by the probability distributions (8) (see also

[2]). If we denote $s(\theta) = P(\{\gamma \in \mathcal{D}_d^m\}) = \sum_{\sigma} \theta_{\sigma(1)} \dots \theta_{\sigma(d)}$ then the rational map

$$p(\theta) = (p_1(\theta), \dots, p_n(\theta)) = \left(\frac{g_1(\theta)}{s(\theta)}, \dots, \frac{g_n(\theta)}{s(\theta)} \right) \quad (11)$$

defines a vector of conditional probabilities where the i -th coordinate p_i represents the probability that the data γ falls into the chamber S_i , given that it was generated by some d -tuple of reactions. In what follows we assume that our data consists of various γ values obtained from different realizations of the system (5) under the generating probability

distribution (8). This setup attempts to mimic the data collection process where multiple independent experiments are conducted, possibly under different external conditions [2].

Consider N different experiments resulting in the set of observed values $\{\gamma_1, \dots, \gamma_N\}$ and the corresponding set of counts of observed values in each chamber $\{u_1, \dots, u_n\}$. The corresponding likelihood function satisfies

$$\mathcal{L}(\theta|\gamma) \propto \prod_{i=1}^n p_i^{u_i}(\theta)$$

with the log-likelihood

$$\ell(\theta|\gamma) = \sum_{i=1}^n u_i \log(p_i(\theta)). \quad (12)$$

It follows that the likelihood and log-likelihood functions are both maximized in $\mathbb{R}^m$ by any $\hat{\theta}_N$ which satisfies $p_i(\hat{\theta}_N) = u_i/N$, where u_i counts the number of γ points that fall into the chamber S_i . The issues of obtaining $\hat{\theta}_N$ in both restricted and unrestricted domains was considered in some detail in [3].

For illustration, in Figure 1 below we consider the set of reactions $R = \{A_1 \rightarrow A_2, A_1 \rightarrow 2A_2, A_1 \rightarrow A_1 + A_2, A_1 \rightarrow 2A_1\}$ in two dimensional species space. The figure shows the partition of the reactions cone in 2D into three chambers S_1, S_2, S_3 , (marked with the Roman numerals I, II, III) each containing hypothesized observed values of γ based on 6 independent experiments.

2.3. Sparsistency

The solution to the set of equations $p_i(\hat{\theta}_N) = u_i/N, i = 1, \dots, n$ may not exist under the domain restrictions on θ_N . However, it always exists, at least asymptotically, if we consider only the $\{0, 1\}^d$ domain. The purpose of this section is to formulate and prove the following simple result.

Theorem 1 (Sparsistency)—Consider the system of equations (7) with N sets of random coefficients where we assume that for each single source sub-system (5) its values of γ come from the probability distribution generated by the d -tuple of reactions

$\sigma_0 = \sigma_0(1) \dots \sigma_0(d) \in \mathcal{D}_d^m$, with the κ_k 's following (8). Then any maximizer of (12), say $\hat{\theta}_N$, satisfies $\hat{\theta}_N \rightarrow \hat{\theta}$, as $N \rightarrow \infty$, almost surely, where $\hat{\theta}$ is the vector that discards the superfluous cones, that is, $\hat{\theta}_{\sigma(1)} \dots \hat{\theta}_{\sigma(d)} = 0$ for all $\sigma \neq \sigma_0$.

Proof—For the proof it suffices to consider a single source network (5), since otherwise we would simply repeat the argument for each cone in (7). Since the permutation $\sigma_0(1), \dots, \sigma_0(d)$ corresponds to the d -tuple of true reactions, the vector $p(\theta)$ associated with the true probability of observations in chamber S_i is

$$p(\theta^0) = (p_1(\theta^0) \dots p_n(\theta^0)) \\ = \left(\frac{\text{vol}(C_{\sigma_0} \cap S_1)}{\text{vol}(C_{\sigma_0})}, \dots, \frac{\text{vol}(C_{\sigma_0} \cap S_n)}{\text{vol}(C_{\sigma_0})} \right)$$

where C_{σ_0} is the cone spanned by the d -tuple $\sigma_0(1), \dots, \sigma_0(d)$. By the consistency properties of the empirical multinomial proportions, as $N \rightarrow \infty$, $u_i/N \rightarrow p_i(\hat{\theta}) = \frac{\text{vol}(C_{\sigma_0} \cap S_i)}{\text{vol}(C_{\sigma_0})}$ with probability one and therefore the asymptotic maximum likelihood estimator or MLE $\hat{\theta}$ (which has no dependence on N) must satisfy $p(\hat{\theta}) = p(\theta^0)$. Since by assumption $s(\hat{\theta}) = 1$ then, due to (11), $p_i(\hat{\theta}) = g_i(\hat{\theta})$ and in view of (10) the asymptotic MLE is the solution to the set of multivariate equations for $i = 1 \dots, n$

$$\sum_{C_\sigma \in \mathcal{R}_d} \left(\frac{\text{vol}(C \cap S_i)}{\text{vol}(C)} \right) \hat{\theta}_{\sigma(1)} \dots \hat{\theta}_{\sigma(d)} = \frac{\text{vol}(C_{\sigma_0} \cap S_i)}{\text{vol}(C_{\sigma_0})}.$$

Summing over i we get

$$\sum_{i=1}^n \sum_{C_\sigma \in \mathcal{R}_d} \left(\frac{\text{vol}(C \cap S_i)}{\text{vol}(C)} \right) \hat{\theta}_{\sigma(1)} \dots \hat{\theta}_{\sigma(d)} = \frac{\text{vol}(C_{\sigma_0})}{\text{vol}(C_{\sigma_0})} = 1$$

or, equivalently, $\sum_{C_\sigma \in \mathcal{R}_d: \sigma \neq \sigma_0} \hat{\theta}_{\sigma(1)} \dots \hat{\theta}_{\sigma(d)} = 0$ which yields the result.

In the theorem above we have assumed that the conic sub-networks are generated by the d -tuples of reactions. However, the SASM method can still be applied when the assumption is violated, although in that case the sparsistency property is not always guaranteed. Also, as discussed in [3], in many practical situations the SASM approach will require a pre-processing step for the network dimension reduction. It is also easy to see that the assumed probability law for the κ coefficients (gamma family) is not particularly relevant for establishing the sparsistency property. Rather, as discussed in the introduction, what is required is that the stochastic law of the rates be empirically identifiable. Finally, let us also note that in the proof above we have tacitly assumed that the values of γ coefficients were known. Typically, this is not the case in practice, however, it is also easy to see from the proof that the theorem's hypothesis holds as long as one may substitute for γ their consistent estimates based on the trajectory values. Recall from Section 1 that the estimates satisfying the consistency requirements were recently discussed in [7, 12].

3. DATA EXAMPLES

In this section for illustration purpose we consider two examples of applying SASM to analyze both linear and non linear biochemical network models. We start with a simple in silico example of a linear network.

3.1. In Silico Example

Consider the following linear biochemical system of two species (X_1, X_2) and two conic subnetworks

with the corresponding mass action ODEs (4) specializing to the following

$$\begin{aligned} \dot{c}_1(t) &= (\kappa_1 - \kappa_3 - \kappa_5)c_1(t) + (\kappa_4 + 2\kappa_6)c_2(t) \\ \dot{c}_2(t) &= (\kappa_5 + 2\kappa_3)c_1(t) + (\kappa_2 - \kappa_4 - \kappa_6)c_2(t). \end{aligned}$$

Reparametrizing in terms of γ gives the decomposition (7), with

$$s=2, c_{(1)}^{v_0}(t)=c_1(t), c_{(2)}^{v_0}(t)=c_2(t), \gamma_1=(\gamma_{11}, \gamma_{21})^\top, \gamma_2=(\gamma_{12}, \gamma_{22})^\top$$

$$\begin{aligned} \dot{c}_1(t) &= \gamma_{11}c_1(t) + \gamma_{12}c_2(t) \\ \dot{c}_2(t) &= \gamma_{21}c_1(t) + \gamma_{22}c_2(t). \end{aligned} \quad (14)$$

To demonstrate the SASM analysis method we generated the set of $N = 15$ data points γ from (14), with the top row of reactions in (13) (the reactions 1,2) being superfluous and the remaining four being true, i.e. with the following parameters

$$\begin{array}{ll} \kappa_1=0 & \kappa_2=0 \\ \kappa_3 \sim \Gamma(2, 2) & \kappa_4 \sim \Gamma(1, 2) \\ \kappa_5 \sim \Gamma(1.2, 2) & \kappa_6 \sim \Gamma(3, 2). \end{array}$$

The γ values generated using the above set of probability distributions are displayed in Figure 2. In order to perform the necessary numerical calculations, the software suite “BioReactor” ([5]) discussed in [3] was used. In particular, the software allowed to compute the MLE $\hat{\theta}_N$ in (12). The results of the SASM analysis are presented in Table I and indicate that the algebraic model has picked up the correct sparsity pattern by discarding the superfluous reaction from each cone.

3.2. Real Data Example

As our second example, let us consider a typical biological network inference problem based on longitudinal data, where a nonlinear biochemical dynamical system is used as a proxy for the model of interactions. Consider the longitudinal RNA-seq data from the zebrafish tissue collected from genetically identical animals at 9 distinct time points after cell specific ablation performed to the rod photoreceptors in the animal retina. Such data is of interest, for instance, when studying the regenerative properties of zebrafish ([1, 8]). In this specific case, two types of animal tissues, under treatment (denoted MTZ) and control (denoted CTL) conditions, were considered, along with a handful of transcription factors (TFs) of interest in the regeneration process signaling. For the purpose of our current example, we

consider the interactions between the heat shock protein transcription factor (*Hsp70*), signal transducer and activator of transcription 3 (*Stat3*), and the known inhibitor of *Stat3*, the suppressor of cytokine signaling 3 (*Socs3b*). For more details on the biological model and further references, see chapter 7 in [6]. For illustration, the data from two trajectories only (one MTZ and one CTL) is presented in Table II. The specific network of interest may be written out in the chemical notation as follows

which, following our convention from Section 1, is equivalent to the mass action ODEs

$$\begin{aligned} \frac{dStat3}{dt} &= \kappa_1 + (\kappa_4 - \kappa_5 - \kappa_6 - \kappa_8)Stat3 - \kappa_7Stat3Soc3b \\ \frac{dSoc3b}{dt} &= \kappa_2 + \kappa_5Stat3 - \kappa_9Soc3b \\ \frac{dHsp70}{dt} &= \kappa_3 + \kappa_6Stat3 - \kappa_{10}Hsp70, \end{aligned}$$

or to the five source system

$$\begin{aligned} \frac{dStat3}{dt} &= \gamma_{10} + \gamma_{11}Stat3 - \gamma_{12}Stat3Soc3b \\ \frac{dSoc3b}{dt} &= \gamma_{20} + \gamma_{21}Stat3 - \gamma_{23}Soc3b \\ \frac{dHsp70}{dt} &= \gamma_{30} + \gamma_{31}Stat3 - \gamma_{34}Hsp70. \end{aligned} \quad (15)$$

From these ODEs one may extract the sub-network of interest associated with the transfer of *Stat3* as presented in Figure 3. Note that this network has an additional reaction $Stat3 \rightarrow \emptyset$ representing a possible role of *Stat3* as an activator for some additional reactions not accounted for in (15). Unlike in our previous (in silico) example, the γ coefficient values in (15) are unknown. However, under the assumption (7) we may use the FPKM data from Table II to estimate their values for each trajectory, following the weighted least squares method discussed in [7, 12]. The resulting set of $\hat{\gamma}_2 = (\hat{\gamma}_{12}, \hat{\gamma}_{22}, \hat{\gamma}_{32})^\top$ estimates for the 8 different empirical trajectories considered in the experiment is presented in Table III.

The results of SASM analysis for the zebrafish data based on the 8 trajectories and the estimated values of γ as given in Table III are summarized in Table IV. By rejecting the additional degradation reaction $Stat3 \rightarrow \emptyset$ as superfluous, the SASM analysis seems to confirm that there is empirical evidence for *Stat3* being a transcription factor for both *Hsp70* and *Socs3b* as well as for its autoregulatory capabilities (since the reaction $Stat3 \rightarrow 2Stat3$ is also inferred as valid). The rejection of the degradation reaction may also indicate that this process occurs on much slower timescale than the one considered in the experiment.

Additional, more targeted experiments would likely be required to study this issue in further detail.

3.3. Software

Let us briefly comment for the interested readers on the software used to carry out calculations in the two numerical examples above. In both cases, in order to calculate the maximizers in (12) we used the software suit 'Bioreactor' developed by one of the authors. The Bioreactor software is currently freely available for use via the web-based interface [5]. The software web site also contains detailed documentation with examples. In addition to the Bioreactor suite, in our biological data analysis we used the RNA-seq preprocessing software TopHat [14] for the reads alignment to the reference genome and Cufflinks [15] for the quantification of the aligned sequencing data.

4. SUMMARY AND CONCLUSIONS

The paper discussed the likelihood-based stoichiometric algebraic statistical model (SASM) applied to biochemical dynamical systems decomposable into disjoint single source subsystems (conic subnetworks). It was shown that, under certain conditions on the generating networks, the model is asymptotically capable of identifying the true species interactions, the property which we dubbed sparsistency. Whereas this result presented is only true asymptotically with the increasing number of observed trajectories, we have also shown with two simple examples that the algebraic model can provide meaningful answers with only the moderate-to-small number of trajectories observed. The potential advantage of the SASM over various reverse engineering paradigms proposed in the literature is its capability of relating the trajectory data to the prescribed geometry of the reaction network. This makes SASM potentially much more robust to the measurement error and the stochastic fluctuations of the experimental system observables, which are two issue of primary concern with biological modeling.

Acknowledgments

The work was partially supported by the US National Science Foundation under grant DMS-1318886 and the US National Cancer Institute under grant R01-CA152158. The authors are grateful to the anonymous referee for his comments.

References

1. Close, Jennie L.; Gumuscu, Burak; Reh, Thomas A. Retinal neurons regulate proliferation of postnatal progenitors and Müller glia in the rat retina via TGF β signaling. *Development*. 2005; 132(13):3015–3026. [PubMed: 15944186]
2. Gheorghe, Craciun; Casian, Pantea; Rempala, Grzegorz A. Algebraic methods for inferring biochemical networks: a maximum likelihood approach. *Comput Biol Chem*. Oct; 2009 3307(5): 361–7. 014.10.1016/j.compbiolchem.2009 [PubMed: 19709932]
3. Gheorghe, Craciun; Jaejik, Kim; Casian, Pantea; Rempala, Grzegorz A. Statistical model for biochemical network inference. *Commun Stat Simul Comput*. Jan; 2013 42(1):121–137.10.1080/03610918.2011.633200 [PubMed: 23125476]
4. Gillespie, Daniel T. A rigorous derivation of the chemical master equation. *Physica A: Statistical Mechanics and its Applications*. 1992; 188(1):404–425.

5. Kim, Jaeyik; Rempala, Grzegorz A. Bioreactor simulation software. 2013. URL <https://neyman.cph.ohio-state.edu/bioreactor.html>
6. Linder, Daniel. PhD thesis. Georgia Regents University; Augusta, GA: Aug. 2013 Penalized least squares and algebraic statistical model for biochemical reaction networks.
7. Linder, Daniel; Rempala, Grzegorz A. Bootstrapping least squares estimates for biochemical reaction networks. Manuscript. Aug.2013
8. Fumitaka, Osakada; Sotaro, Ooto; Tadamichi, Akagi; Michiko, Mandai; Akinori, Akaike; Takahashi, Masayo. Wnt signaling promotes regeneration in the retina of adult mammals. *J Neurosci.* Apr; 2007 27(15):4210–9.10.1523/jneurosci.4193-06.2007 [PubMed: 17428999]
9. Perez, Omar D.; Krutzik, Peter O.; Nolan, Garry P. Flow cytometric analysis of kinase signaling cascades. *Methods Mol Biol.* 2004; 263:67–94.10.1385/1-59259-773-4:067 [PubMed: 14976361]
10. Qin Z, Raymond PA. Microarray-based gene profiling analysis of muller glia derived retinal stem cells in light-derived retinas from adult zebrafish. *Methods Mol Biol.* 2012; (884):255–61. [PubMed: 22688712]
11. Natasa, Rajicic; Joseph, Cuschieri; Finkelstein, Dianne M.; Miller-Graziano, Carol L.; Hayden, Douglas; Moldawer, Lyle L.; Moore, Ernest; O’Keefe, Grant; Pelik, Kimberly; Warren, H Shaw; Schoenfeld, David A.; Inflammation and the Host Response to Injury Large Scale Collaborative Research Program. Identification and interpretation of longitudinal gene expression changes in trauma. *PLoS One.* 2010; 5(12):e14380.10.1371/journal.pone.0014380 [PubMed: 21187951]
12. Rempala, Grzegorz A. Least squares estimation in stochastic biochemical networks. *Bull Math Biol.* Aug; 2012 74(8):1938–55.10.1007/s11538-012-9744-y [PubMed: 22752340]
13. Stanescu, Dan; Chen-Charpentier, Benito M. Random coefficient differential equation models for bacterial growth. *Mathematical and Computer Modelling.* 2009; 50(5):885–895.
14. Cole, Trapnell; Lior, Pachter; Salzberg, Steven L. Tophat: discovering splice junctions with RNA-Seq. *Bioinformatics.* May; 2009 25(9):1105–11.10.1093/bioinformatics/btp120 [PubMed: 19289445]
15. Cole, Trapnell; Hendrickson, David G.; Sauvageau, Martin; Goff, Loyal; Rinn, John L.; Pachter, Lior. Differential analysis of gene regulation at transcript resolution with RNA-Seq. *Nat Biotechnol.* Jan; 2013 31(1):46–53.10.1038/nbt.2450 [PubMed: 23222703]

FIG. 1. Stoichiometric geometry in 2D

The location of 6 γ data points from independent experiment for the conic network $R = \{A_1 \rightarrow A_2, A_1 \rightarrow 2A_2, A_1 \rightarrow A_1 + A_2, A_1 \rightarrow 2A_1\}$.

FIG. 2. The γ coefficients generated in silico

Plot of the γ values of $N = 15$ data points generated from the distributions of κ_3 , κ_4 , κ_5 , and κ_6 ($\kappa_1 = \kappa_2 = 0$).

FIG. 3. Stat3 conic network

The conic network of TFs activations via Stat3 with an added reaction $Stat3 \rightarrow \emptyset$ representing other hypothetical activations not accounted for in (15).

TABLE I**SASM analysis results**

The results of the SASM analysis of the system (14) with $N = 15$ in silica data points performed with the help of the Bioreactor software. The results are consistent with the true network patterns.

Cone	Reaction	$\hat{\theta}_{ij}$
1	$X_1 \rightarrow X_2$	1
1	$X_1 \rightarrow 2X_2$	1
1	$X_1 \rightarrow 2X_1$	0
2	$X_2 \rightarrow X_1$	1
2	$X_2 \rightarrow 2X_1$	1
2	$X_2 \rightarrow 2X_2$	0

TABLE II

Zebrafish time-course next-gen-seq data (in FPKM)

An example of the trajectory data from the longitudinal zebrafish photoreceptor study. The software TopHat [14] was used for the reads alignment to the reference zebrafish genome, Zv9, and the software Cufflinks [15] was used to obtain the expression measure of fragments per kilobase of transcript per million mapped reads (FPKM), which may be treated as a proxy for the required RNA concentration.

Time (hrs)	0	1	2	4	8
Stat3 CTL	16.33	5.38	5.44	4.69	3.96
Stat3 MTZ	16.33	5.49	8.42	4.47	3.31
Soes3b CTL	2.86	4.95	4.53	4.61	6.17
Soes3b MTZ	2.86	4.20	4.38	6.47	13.05
Hsp70 CTL	20.36	28.84	20.69	15.76	15.46
Hsp70 MTZ	20.36	25.10	20.44	16.05	22.41

Time (hrs)	8	16	24	48	72
Stat3 CTL	3.96	2.36	4.96	8.58	11.80
Stat3 MTZ	3.31	15.42	27.81	16.58	12.99
Soes3b CTL	6.17	6.73	4.41	3.27	3.90
Soes3b MTZ	13.05	50.27	33.20	10.56	6.81
Hsp70 CTL	15.46	11.66	17.38	15.31	15.16
Hsp70 MTZ	22.41	13.41	14.65	17.24	15.71

TABLE III**Coefficient estimates**

The trajectory-based $\hat{\gamma}_2$ estimates in the conic subnetwork in Figure 3. The values of $\hat{\gamma}_2$ are calculated using the least squares method discussed recently in [7, 12].

Traj No	$\hat{\gamma}_{11}$	$\hat{\gamma}_{21}$	$\hat{\gamma}_{31}$
1	7.11	0.35	2.48
2	15.13	0.31	1.48
3	8.85	3.79	1.55
4	1.69	1.31	1.83
5	16.81	0.14	4.62
6	3.31	2.16	2.56
7	2.52	1.75	2.37
8	27.70	0.31	0.16

TABLE IV
SASM analysis results for Stat3 network

The final result of the SASM analysis for the conic network in Figure 3, as given by the Bioreactor software [5].

Reaction	Prob (θ_k)
$Stat3 \rightarrow 2Stat3$	1
$Stat3 \rightarrow Socs3b$	1
$Stat3 \rightarrow Hsp70$	1
$Stat3 \rightarrow \emptyset$	0