Algebraic Methods for Inferring Biochemical Networks: a Maximum Likelihood Approach

Gheorghe Craciun*, Casian Pantea†, Grzegorz A. Rempala‡

February 3, 2009

Abstract

We present a novel method for identifying a biochemical reaction network based on multiple sets of estimated reaction rates in the corresponding reaction rate equations arriving from various (possibly different) experiments. The current method, unlike some of the graphical approaches proposed in the literature, uses the values of the experimental measurements only relative to the geometry of the biochemical reactions under the assumption that the underlying reaction network is the same for all the experiments. The proposed approach utilizes algebraic statistical methods in order to parametrize the set of possible reactions so as to identify the most likely network structure, and is easily scalable to very complicated biochemical systems involving a large number of species and reactions. The method is illustrated with a numerical example of a hypothetical network arising form a "mass transfer"-type model.

Keywords: Biochemical reaction network, law of mass action, algebraic statistical model, polyhedral geometry.

2000 AMS Subject Classification: 92C40, 92C45, 52B70, 62F

1 Introduction

In modern biological research, it is very common to collect detailed information on time-dependent chemical concentration data for large networks of biochemical reactions (see survey papers [3, 12]). Often, the main purpose of collecting such data is to identify the exact structure of a network of chemical reactions for which the identity of the chemical species present in the network is known but *a priori* no information is available on the species interactions (see e.g., [13]). The problem is of interest both in

*Department of Mathematics and Department of Biomolecular Chemistry, University of Wisconsin-Madison. Email: craciun@math.wisc.edu, Phone: (608)265-3391, Fax: (608)263-8891.

†Department of Mathematics, University of Wisconsin-Madison. E-mail: pantea@math.wisc.edu.

‡Department of Biostatistics, Medical College of Georgia, Augusta, GA 30912. E-mail: grempala@mcg.edu.

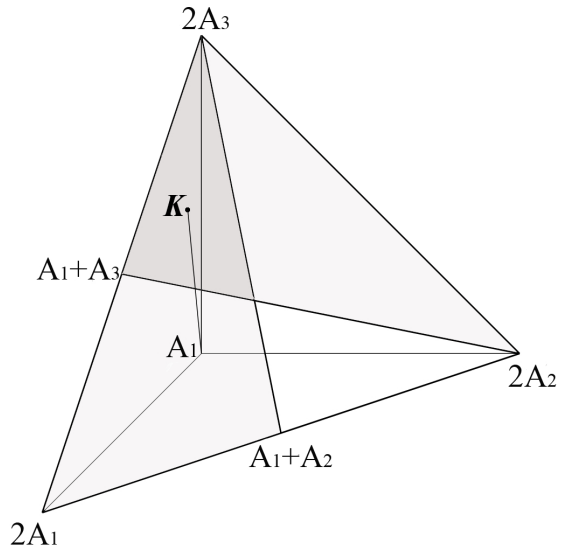


Figure 1: Identifiability of reaction networks given experimental data. Note that, for a deterministic mass-action model, even if we can estimate the vector K of parameter values with great accuracy, we cannot determine if the "correct" reaction network is $\{A_1 \to 2A_1, A_1 \to A_1 + A_2, A_1 \to 2A_3\}$ or $\{A_1 \to 2A_2, A_1 \to A_1 + A_3, A_1 \to 2A_3\}$, because K belongs to the span of either one of these networks. However, if instead a single point K one has available a set $\mathcal{D} = \{K_i, i = 1 \ldots, k\}$, interpreted as a result of random selection of parameter rate values according to some probability law, then the spanning cone of the data points may be used to identify the sets of reactions that "best explain" the data.

the setting of classical theoretical chemistry, as well as, more recently, in the context of molecular and systems biology problems and as such has received a lot of attention in the literature over last several decades as evidenced by multiple papers devoted to the topic [1, 7, 8, 9, 11, 18, 19, 20, 21].

In general, two very different reaction networks might generate identical mass-action dynamical system models, making it impossible to discriminate between them, even if one is given experimental data of perfect accuracy and unlimited temporal resolution. Sometimes this *lack of uniqueness* is referred to as the "fundamental dogma of chemical kinetics", although it is actually not a well known fact in the biochemistry or chemical engineering communities [3, 4, 5]. Necessary and sufficient conditions for two reaction networks to give rise to the same *deterministic* dynamical system model (i.e., the same *reaction rate equations*) are described in [2], where the problem of identifiability of reaction networks given high accuracy data was analyzed in detail. The key observation is that, if we think of reactions as vectors, it is possible for different sets of such vectors to span the same positive cones, or at least to span positive cones that have nonempty intersection (see Figure 1 for an example).

On the other hand, it is often the case that experimental measurements for the study of a specific reaction network or pathway are being collected under many different

experimental conditions, which affect the values of reaction rate parameters. Almost always, the reactions of interest are not "elementary reactions", for which the reaction rates parameters must be constant, but they are so called "overall reactions" that summarize several elementary reaction steps. In that case the reaction rates parameters may reflect the concentrations of biochemical species which have not been included explicitly in the model. In such circumstances the reaction rate parameters are *not* constant, but rather depend on specific experimental conditions, such as concentrations of enzymes and other intermediate species. Therefore, the estimated vector of reaction rate parameters will not be the same for all experimental conditions, but each specific experimental setting will give rise to one such vector of parameters. However, the set of all these vectors should span a specific cone, whose extreme rays should identify exactly the set of reactions that gave rise to the data.

The purpose of the current paper is to propose a statistical method based on the above geometric considerations, which allows one to take advantage of the inherent stochasticity in the data, in order to determine the *unique* reaction network that can best account for the results of *all* the available experiments pooled together. The idea is related to the notion of an *algebraic statistical model* (as described in [14] Chapter 1), and relies on mapping the estimated reaction parameters into an appropriate convex region of the span of reaction vectors of a network, using the underlying geometry to identify the reactions which are most likely to span that region. As shown below, this approach reduces the network identification problem to a statistical inference problem for the parameters of a multinomial distribution, which may then be solved for instance using the classical likelihood methods.

2 Maximum Likelihood Inference for a Biochemical Reaction Network

In this section we develop a formal way of inferring a most likely subnetwork of a given conic network (i.e., network represented by a cone like the one in Figure 1) of the minimal spanning dimension. For the inference purpose, in the network of m reactions we assume that the empirical data $\mathcal{D} = \{K_i, i = 1 \ldots, k\} \subset R^d$ is available in the form of (multiple) estimates of the parameters of the system of differential equations corresponding to a hypothesized biochemical network. As illustrated in Figure 1, such networks are in general "unidentifiable" in the sense that different chemical reaction networks may give rise to the same system of differential equations. However, in the stochastic or "statistical" sense it is possible to identify the "most likely" (i.e., maximizing the appropriate likelihood function) network as indicated by the data $\mathcal{D}$.

2.1 Multinomial model

Consider d species, and m possible reactions with reaction vectors $R = \{R_1, \ldots, R_m\} \subset \mathbb{R}^d$ among the species. (For more details about how each reaction generates a reaction vector see [2].) Let $\mathcal{R}_d$ denote the collection of all $\binom{m}{d}$ positive cones spanned by subsets of d reactions in R.

Denote by $cone(R)$ the positive cone generated by the reaction vectors in R. Let

S be the partition of $cone(R)$ obtained by all possible intersections of non-degenerate cones in $\mathcal{R}_d$. Suppose S contains n full-dimensional regions $S_1, \ldots, S_n$; throughout we shall refer to these regions as *building blocks*, and to n as *the number of building blocks*.

Let Δ_{m-1} be a probability simplex in $\mathbb{R}^m$ and let $\theta \in \Delta_{m-1}$ be a vector of probabilities associated with the reactions that give rise to R. We assume that these m reactions have the same source complex (i.e., form a conic network), since, as explained in [2], the identifiability of a network can be addressed one source complex at a time. Define the polynomial map

$$g : \Delta_{m-1} \to \mathbb{R}^n$$

where

$$g_i(\theta) = \sum_{C=cone(R_{\sigma(1)},\ldots,R_{\sigma(d)})\in\mathcal{R}_d} \frac{vol(C \cap S_i)}{vol(C)} \theta_{\sigma(1)} \cdots \theta_{\sigma(d)} \tag{1}$$

for $i = 1, \ldots, n$.

We take[1] $\frac{vol(C\cap S_i)}{vol(C)} = 0$ if $vol(C) = 0$. Define $s(\theta) = \sum_\sigma \theta_{\sigma(1)} \cdots \theta_{\sigma(d)}$ and

$$p(\theta) = (p_1(\theta) \ldots, p_n(\theta)) = (g_1(\theta)/s(\theta), \ldots, g_n(\theta)/s(\theta)). \tag{2}$$

In this setting $p \in \mathbb{R}^n$ is our statistical model for the data, after we substitute $\theta_m = 1 - \sum_{j=1}^{m-1} \theta_j$. Note that we may interpret the monomials $\theta_{\sigma(1)} \cdots \theta_{\sigma(d)}$ in (1) as the probabilities of a given data point being generated by the d-tuple of reactions $\sigma(1), \ldots, \sigma(d)$. With this interpretation the coordinate p_i of the map p in (2) is simply the conditional probability that the data point is observed in S_i given that it was generated by a d-tuple of reactions. Note that the map p is rational but, as we shall see below, the model may be re-parametrized into an equivalent one involving only the multilinear map (1).

Let u_i denote the number of data points in S_i. The log-likelihood function corresponding to a given data allocation is

$$l(\theta) = \sum_{i=1}^{n} u_i \log p_i(\theta). \tag{3}$$

Our inference problem is to find

$$\hat{\theta} = \text{argmax}_\theta l(\theta) \qquad \text{subject to} \quad \sum_{i=1}^{m} \theta_i = 1 \quad \text{and} \quad \theta_i \geq 0. \tag{4}$$

Example 2.1. *Consider the two reaction networks described in Figure 1. The model has $d = 3$ species and a total of $m = 5$ possible reactions $R = \{R_1, \ldots, R_5\} = \{A_1 \to 2A_1, A_1 \to A_1 + A_2, A_1 \to 2A_3, A_1 \to 2A_2, A_1 \to A_1 + A_3\}$. In this case there are $n = 5$ building blocks $S_1, \ldots, S_5$ defined by the intersections of all non-trivial reaction cones generated by reaction triples, illustrated in Figure 2.*

[1] In general, it may be beneficial to consider various measures $vol(\cdot)$ which are absolutely continuous w.r.t. the usual Lebesque measure. For instance in Section 3 we describe an example where this measure is defined via gamma densities.

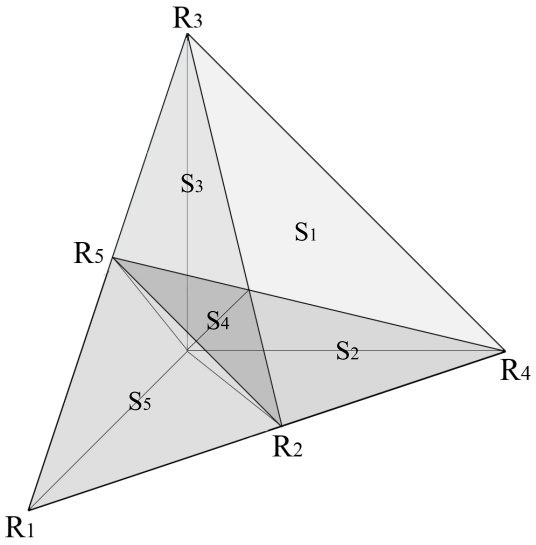


Figure 2: Configuration of the five building blocks for the possible reactions $R = \{R_1, \ldots, R_5\} = \{A_1 \to 2A_1, A_1 \to A_1 + A_2, A_1 \to 2A_3, A_1 \to 2A_2, A_1 \to A_1 + A_3\}$ of example 2.1.

Thus denoting $C_{jkl} = cone(R_j, R_k, R_l)$ *for any triple* $\{j,k,l\} \in \{1,\ldots,5\}$ *we have*

$$\begin{aligned}
S_1 =& C_{134} \cap C_{234} \cap C_{345} \\
S_2 =& C_{134} \cap C_{145} \cap C_{234} \cap C_{245} \\
S_3 =& C_{123} \cap C_{134} \cap C_{235} \cap C_{345} \\
S_4 =& C_{123} \cap C_{134} \cap C_{145} \cap C_{235} \cap C_{245} \\
S_5 =& C_{123} \cap C_{125} \cap C_{134} \cap C_{145}.
\end{aligned}$$

Note that the cones C_{124} *and* C_{135} *are degenerate and are not involved in the definitions of the* S_i*'s. Denoting further* $v_{jkl}^{(i)} = vol(C_{jkl} \cap S_i)/vol(C_{jkl})$ *for any triple* $\{j,k,l\} \in \{1,\ldots,5\}$*, we see that the the map* (1) *becomes*

$$\begin{aligned}
g_1(\theta) =& v_{134}^{(1)}\theta_1\theta_3\theta_4 + v_{234}^{(1)}\theta_2\theta_3\theta_4 + v_{345}^{(1)}\theta_3\theta_4\theta_5 \\
g_2(\theta) =& v_{134}^{(2)}\theta_1\theta_3\theta_4 + v_{145}^{(2)}\theta_1\theta_4\theta_5 + v_{234}^{(2)}\theta_2\theta_3\theta_4 + v_{245}^{(2)}\theta_2\theta_4\theta_5 \\
g_3(\theta) =& v_{123}^{(3)}\theta_1\theta_2\theta_3 + v_{134}^{(3)}\theta_1\theta_3\theta_4 + v_{235}^{(3)}\theta_2\theta_3\theta_5 \\
g_4(\theta) =& v_{123}^{(4)}\theta_1\theta_2\theta_3 + v_{134}^{(4)}\theta_1\theta_3\theta_4 + v_{145}^{(4)}\theta_1\theta_4\theta_5 + v_{235}^{(4)}\theta_2\theta_3\theta_5 + v_{245}^{(4)}\theta_2\theta_4\theta_5 \\
g_5(\theta) =& v_{123}^{(5)}\theta_1\theta_2\theta_3 + \theta_1\theta_2\theta_5 + v_{134}^{(5)}\theta_1\theta_3\theta_4 + v_{145}^{(5)}\theta_1\theta_4\theta_5,
\end{aligned}$$

where the coefficients satisfy $\sum_i v_{jkl}^{(i)} = 1$ *for any triple* $\{j,k,l\}$ *appearing on the right-*

hand-side in the formulas above. The rational map (2) *is therefore given by*

$$p = \frac{g}{\sum_{jkl} \theta_j \theta_k \theta_l}$$

where the sum in the denominator extends over all distinct triples $\{j,k,l\}$ *excluding* $\{1,2,4\}$ *and* $\{1,3,5\}$*, i.e., the ones corresponding to degenerate cones.*

2.2 Multilinear representation

The model representation via a rational map (2) may be equivalently described in terms of a simpler polynomial map (1) as follows. Let us substitute $\tilde{\theta}_i = \theta_i s^{-1/d}$ for $i = 1, \dots m$ and define

$$\tilde{g}_i(\tilde{\theta}) = p_i(\theta).$$

Note that $\tilde{g}_i : \mathbb{R}^m_{>0} \to \mathbb{R}^n$ and $l(\tilde{\theta}) = l(\theta)$. Thus we may consider a following more convenient version of (4). Find

$$\begin{aligned} &\hat{\theta} = \operatorname{argmax}_{\tilde{\theta}}\ l(\tilde{\theta}) \\ &\text{subject to} \quad \sum_{\sigma} \tilde{\theta}_{\sigma(1)} \cdots \tilde{\theta}_{\sigma(d)} = \sum_i \tilde{g}_i(\tilde{\theta}) = \sum_{i=1}^n p_i(\theta) = 1, \quad \forall_i\ \tilde{\theta}_i \geq 0. \end{aligned} \tag{$*$}$$

Consider a fixed d-tuple of reactions (say, σ_1) and in the formulas for $\tilde{g}_i$ $(i = 1 \dots, n)$ substitute $\tilde{\theta}_{\sigma_1(1)} \cdots \tilde{\theta}_{\sigma_1(d)} = 1 - \sum_{\sigma \neq \sigma_1} \tilde{\theta}_{\sigma(1)} \cdots \tilde{\theta}_{\sigma(d)}$. Note that the resulting algebraic statistical map is multilinear i.e, linear in one parameter $\tilde{\theta}_k$ when all others are fixed. For instance, as a function of $\tilde{\theta}_1$ we have

$$p_i(\tilde{\theta}_1 | \cdot) = a_i \tilde{\theta}_1 + b_i \quad i = 1 \dots, n$$

where $\sum_i a_i = 0$ and $\sum_i b_i = 1$ and a_i, b_i are given in terms of $\tilde{\theta}_l$ for $l > 2$.

By Varchenko's theorem (see, [14] chapter 1) the conditional, one dimensional version of problem $(*)$ may be now solved iteratively for each $p_i(\tilde{\theta}_1|\cdot)$, $i = 1, \dots, n$ by finding a unique root of the score equations in the regions bounded by the ratios $-b_i/a_i$.

Maximization algorithm. Due to the conditional convexity of the one dimensional problems the above considerations suggest that the following algorithm for (local) maximization of $l(\tilde{\theta})$ should be valid (cf. also [14], Example 1.7, page 11):

Algorithm 2.1.

1. *Pick initial vectors* $\tilde{\theta}$ *and* $\tilde{\theta}_{old} \in \mathbb{R}^m$.
2. *While* $|l(\tilde{\theta}) - l(\tilde{\theta}_{old})| > \epsilon$
   - $\tilde{\theta}_{old} \leftarrow \tilde{\theta}$
   - *for k=1 to m do*
     - *compute* a_i, b_i *(as functions of* $\tilde{\theta}_j, j \neq k$*)*
     - *identify the bounded interval as determined by Varchenko's fromula which is statistically meaningful (there is only one).*

- *use a simple hill-climbing algorithm to find an optimal* $\tilde{\theta}_k^{opt}$ *in that interval*
- *update* $\tilde{\theta}_k \leftarrow \tilde{\theta}_k^{opt}$

3. Recover θ *from* $\tilde{\theta}$ *by taking* $\theta_k = \tilde{\theta}_k / \sum_i \tilde{\theta}_i$.

The advantage of the algorithm above is that it reduces a potentially very complicated multivariate optimization problem in which d and m are large to iteratively solving of a simple, univariate one. The disadvantage is that due to its dimension-iterative character the algorithm is seen to be slow and for smaller networks perhaps less efficient than some off-the-shelf optimization algorithms available in commercial software (e.g., some modified hill-climbing methods with random restarts). For that reason in our numerical example below we used the standard Matlab optimization package rather than Alg. 2.1.

In the reminder of the paper we revert to the notation of Section 1 and the original problem (4). Based on $(*)$ in this section we may thus extend map g to $\mathbb{R}^m_{>0}$, take $s(\theta) = 1$ in (3) and re-cast the original likelihood maximization problem (4) as

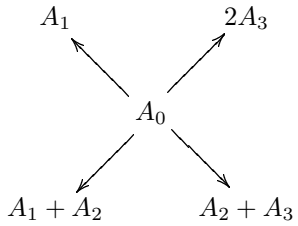
$$\hat{\theta} = \text{argmax}_\theta \sum_i u_i \log g_i(\theta) \qquad \text{subject to} \quad \sum_\sigma \theta_{\sigma(1)} \cdots \theta_{\sigma(d)} = 1 \quad \text{and} \quad \theta_i \geq 0 \quad (4')$$

where the g_i's are given by (1).

3 Simulated Numerical Example

In this section we illustrate the ideas discussed above by analyzing a specific numerical example in detail.

If we have d chemical species and data of the form $\mathcal{D} = \{K_i, i = 1 \ldots, k\}$, then we would hope that the statistical algorithm described above should recover the most likely d reactions out of a given list of $m \geq d$ possible reactions, by finding the maximizing vector $\hat{\theta}$ of the corresponding log-likelihood function. In what follows the setup of the problem is that of $(4')$. To this end, consider the following four-dimensional example:



(5)

where A_i, $i = 0, 1, 2, 3$, denote four chemical species. We shall use the above reaction network to simulate "experimentally measured" data and to test the performance of our method outlined in Section 2. To this end we shall augment the above network by including one or more "incorrect" reactions, and shall check whether our likelihood-based algorithm $(4')$ is able to identify the original "correct" set of four reactions.

Parameter	True Values	Estimated Values	Estimators SEs
γ_0	(-6.259, -4.097,-3.369)	(-5.974, -4.134, -3.63)	(0.210, 0.148, 0.125)
γ_1	(3.310, 2.205, 2.451)	(3.012, 2.254, 2.75)	(0.148, 0.109, 0.106)
γ_2	(1.281, 0.966, 1.805)	(1.199, 1.017, 1.93)	(0.107, 0.082, 0.093)
γ_3	(4.945, 3.154, 1.771)	(5.026, 3.092, 1.704)	(0.190, 0.124, 0.088)

Table 1: Three sets of parameters $(\gamma_0, \gamma_1, \gamma_2, \gamma_3)$ of system (6) corresponding to the trajectories depicted in Figure 3 along with their estimated values (obtained via least-squares fitting) and standard errors of the estimates.

Data generation. Note that the (deterministic) dynamics of the chemical reaction network (5) is governed by linear differential equations of the form

$$dA_i/dt = \gamma_i A_0 \qquad i = 0, \ldots, 3. \tag{6}$$

where the parameters γ_i, $i = 0 \ldots 3$ are linear combinations of the rate constants, given by the stoichiometry of the true network. In our example each data point $K_i \in \mathcal{D}$ $(i = 1, \ldots, k)$ was generated by estimating the set of parameters (γ_i) of the true reaction network (5). For each one of the k data points the set of parameters (γ_i) was obtained from rate constants drawn independently from a gamma distribution $G(\alpha, \lambda)$ with parameters $\alpha = 1.5$ and $\lambda = 1$. In order to identify the coordinates of the points in $\mathcal{D}$, the estimated parameters $\hat{\gamma}_i$, $i = 0, 1, 2, 3$, were calculated each time by fitting the trajectories (6) to the time series data points generated from the stochastic process tracing (6) (see [6]). The Gillespie algorithm (see [15]) was used to generate the 20 equally-spaced values of the trajectory of random process on the interval $(0, 1)$ with the fixed initial condition. An example of three random trajectories with independently generated reaction constants values is given in Figure 3. These three trajectories would give rise to three independently estimated sets of values (γ_i) and consequently to three data points $K_i \in \mathcal{D}$. The fitting was based on the least-squares criterion which is statistically justified for estimation purpose of (γ_i) in this particular case by an appropriate central limit theorem (cf. e.g., [6] chapter 11).

We view the single resulting $(\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2, \hat{\gamma}_3)$ as coordinates of a point K in the species coordinate system $[A_0, A_1, A_2, A_3]$. This representation of data points is not related to a choice of the reactions; note, however, that if the estimation error is sufficiently small[2], each data point lies inside the open convex cone generated by the true reactions (5). As shown in [2] the coordinates of K_i in the basis given by the reaction vectors in (5) are precisely the estimates of the true rate constants.

The data set $\mathcal{D} = \{K_i, i = 1 \ldots, k\}$ used in the simulation described above was based on multiple batches of $k = 10$ data points. The first three data points from the

[2]Here we assume tacitly that the estimation error is sufficiently small and that the statistical estimation procedure is consistent. It turns out this is typically the case in the settings similar to our simulated example, but the discussion of the precise conditions under which this is true in real experimental settings goes beyond the scope of our present discussion. For our current example a brief inspection of the Table 3 indicates a reasonably good agreement between the estimates and the true values of the parameters $(\gamma_0, \gamma_1, \gamma_2, \gamma_3)$ both in terms of actual values as well as the corresponding SE's.

first batch are summarized in Table 3.

In order to test our method, let us first add one incorrect reaction, $A_0 \to A_2$, and from this point on suppose we have no *a priori* knowledge of the true chemistry; therefore, the five possible reactions are as shown in (7).

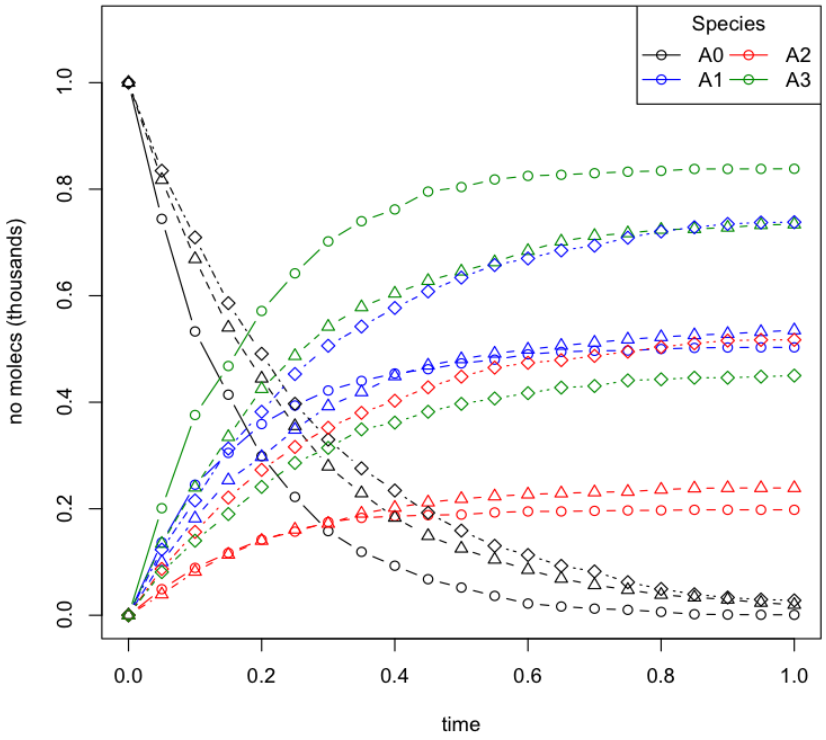
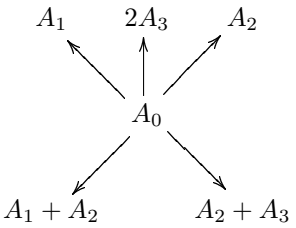


Figure 3: An example of generation of the data points $K_i \in \mathcal{D}$ for $i = 1, 2, 3$ via a two-step process of simulation and estimation. Three stochastic trajectories of the reaction network (5) were simulated via Gillespie algorithm with propensity (reaction) constants drawn randomly according to gamma $G(1.5, 1)$ distribution. The trajectories values at the data collection points are marked at 20 equally-spaced time-points from 0 to 1. The data from the set of trajectories was used in order to estimate the coordinates of $K_i = (\gamma_0, \gamma_1, \gamma_2, \gamma_3)$, $i = 1, 2, 3$. The numerical values of the coordinates corresponding to the given trajectories along with their least-squares estimates are presented in Table 1.



(7)

Later in this section we also consider the case where we add not just one, but several incorrect reactions.

Calculation of the log-likelihood function. In order to obtain an estimate $\hat{\theta}$ via $(4')$ one needs to be able to evaluate the map (1), i.e., in addition to the data counts vector $u \in \mathbb{R}^n$ in (3) one also needs to know the values of the coefficients of the polynomial map. Whereas the calculation of the exact values is difficult for $d > 2$, one may typically resort to Monte-Carlo approximations (see, e.g., [10]). In our current example, for a non-degenerate (i.e., 4-dimensional) cone C, we have computed the approximate relative volumes $vol(C \cap S_i)/vol(C)$ using the following Monte Carlo method. For each cone C we generated $N = 2000$ points inside C with the corresponding conical coordinates randomly drawn from the four independent gamma$G(1.5, 1)$ random variable and then counted the proportion of the total points falling into $C \cap S_i$ i.e., used the approximation

$$\frac{vol(C \cap S_i)}{vol(C)} \approx (\#\text{points in } C \cap S_i)/N \quad i = 1 \ldots, n.$$

With the coefficient values determined as above, the coordinate polynomial maps g_i in (1) were easily calculated now by identifying the cones that contained the appropriate building block regions S_i.

Visualization of the chemical network. The geometry in (7) can be visualized in the 3-dimensional subspace $\mathcal{W} \subset \mathbb{R}^n$ generated by $\{A_1, A_2, A_3\}$. This follows as all the reaction targets are in this subspace, and we can understand the configuration of relevant four-dimensional cones by looking at their intersections with $\mathcal{W}$. Each four-dimensional cone with vertex X_0 intersects $\mathcal{W}$ along a tetrahedron. The intersections of all these tetrahedra cut out the building blocks corresponding to our example (7), as illustrated in Figures 4 and 3. There are five vertices labeled by numbers corresponding to the five target reactions in (7); they form a six-faced convex polyhedron $\mathcal{P}$. Let C be the intersection of line passing through points 1 and 2, denoted (12), with the plane (345). Then all building blocks are tetrahedra with a vertex at C and the opposite face being one of the six faces of the polyhedron $\mathcal{P}$. For example, the building block $(C245)$ is depicted in Figure 4.

Not surprisingly, the 10 data points generated in our example were distributed among the building blocks that compose the tetrahedron (1234) corresponding to the true reactions; 6 data points fell inside the building block $(C234)$ and 4 inside $(C134)$. The log-likelihood function was found in this case as

$$l(\theta) = 6\log(.706 \cdot \theta_1\theta_2\theta_3\theta_4 + .35 \cdot \theta_2\theta_3\theta_4\theta_5) + 4\log(.294 \cdot \theta_1\theta_2\theta_3\theta_4 + .339 \cdot \theta_1\theta_3\theta_4\theta_5).$$

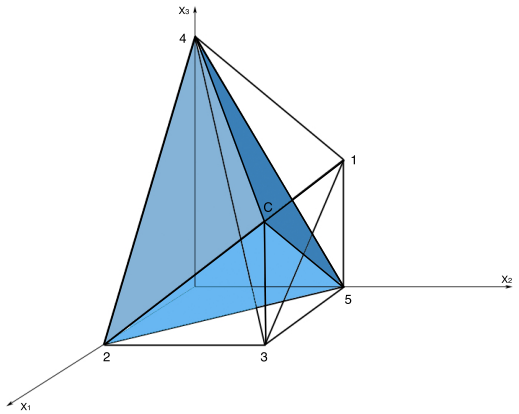


Figure 4: Geometry of building blocks for reaction network (7).

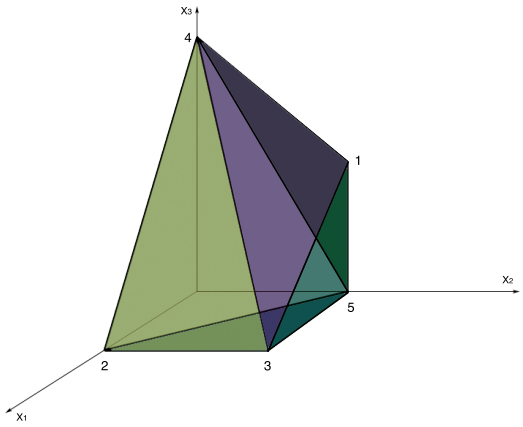


Figure 5: Faces of polyhedron corresponding to reaction network (7).

Maximization of log-likelihood. In order to maximize $l(\theta)$ or, equivalently, to minimize $-l(\theta)$, rather than using the conditional Algorithm 2.1 we used the Matlab function `fmincon` for constrained optimization. As in (4'), the constraint is given by the condition $s(\theta) = \sum_\sigma \theta_{\sigma(1)} \cdots \theta_{\sigma(4)} = 1$ and comes from the fact that there are 4 reactions in the true network. This constraint also assures that $\tilde{g}$ maps into the probability simplex Δ_{n-1}, i.e., defines an algebraic variety (polynomial map) which corresponds to a valid tstatistical model.

The optimization is repeated 2^{m-1} times (i.e., 16 times for the example for network (7)) with random initial conditions satisfying the constrain. A list of (local) minima was created and entries were merged if they were sufficiently close. The point θ that realized the smallest local minimum was reported together with the percentage of time the algorithm ended up at that particular point (success rate). The output for example (7) given by the customized Matlab function was

```
Minimum of negative log-likelihood:  6.94
Theta:
1           1           1           1           0.
Hits:  16 out of 16, 100%.
```

As we may see from these results, in the notation of Figures 4 and 3 the algorithm identified the true reactions (targets) 1, 2, 3 and 4 and discarded the incorrect reaction 5.

More numerical comparisons. We also ran example (5) with, respectively, two, three, four and five incorrect reactions added to the set of four correct ones. The true network was always identified and the success rate (percentage of correct hits for various random initial guess) was high. The results of these experiments are summarized in Table 2. Aditionally, in order to investigate the robustness of our procedure against

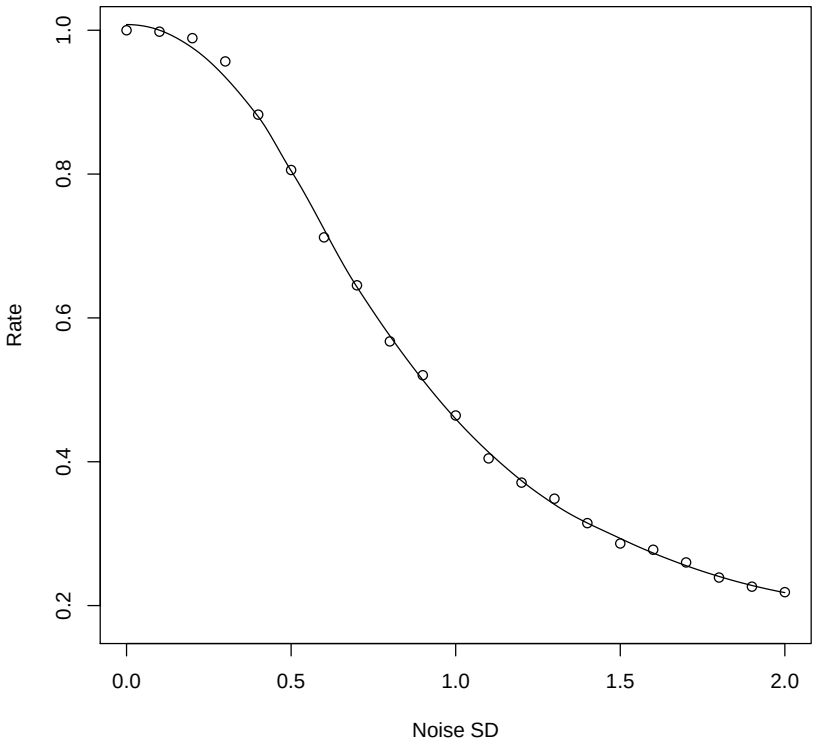


Figure 6: The rate of recovery of the correct reactions out of the network (7) as a function of the size of the Gaussian noise added to the least-squares estimates of the (γ_i) parameters. The solid graph is a smoothed representation of the empirical values (marked as circles) based on 3000 batches of $k = 10$ sets of estimates. The recovery rate is over 90% for the Gaussian noise with 0.5 SD and about 50% for 1 SD.

the lack of precision in the estimates of the γ's, we have added the zero-mean Gaussian term with varying variance to the multiple batches (3000) of $k = 10$ sets of estimators $\hat{\gamma}$ in the network (7) with one erroneous reaction. The estimated probability of assigning the lowest probability to the extraneous reaction vector as a function of increased noise added to the mean-squared estimated values $\hat{\gamma}$ is presented in Figure 6. As seen from the figure with small to moderate random noise added to the values of the reaction constants the likelihood method is still able to recover the correct set of reactions at remarkably high rate. Not surprisingly, for very noisy data (above one SD) the recovery rate decreases significantly.

4 Summary and Discussion

We have proposed herein a statistical method for inferring a biochemical reaction network given several sets of data that originate from "noisy" versions of the reaction rate equations associated with the network. As illustrated in the earlier work of some of the authors [2], in the usual deterministic sense such networks are in general unidentifiable, i.e., different chemical reaction networks may give rise to exactly the same reaction rate equations. In practice, the matters are further complicated since the coefficients of the reaction rate equations are estimated from available experimental data, and hence are subject to measurement error and, moreover, their actual values may differ at different experimental conditions, i.e. at different data points. The statistical approach described here is largely unaffected by these problems, as it only relies on the geometry of the network relative to the data distribution, in order to identify the sets of most likely reactions. Hence, the method takes advantage of the algebraic and geometric representation of the network rather than merely the observed experimental values of the network species, as is commonly the case in network inference models based on graphical methods, like e.g. Bayesian or probabilistic boolean networks (cf. [16]). Still, in order to use the proposed multinomial parametrization of a biochemical network, the method does require a valid way of mapping the experimentally estimated rate coefficients into the networks' appropriate convex regions, and with very large measurement errors is likely to perform poorly as illustrated in Figure 6. On the other hand, precisely because of the need for the experimental data mapping, the method has a very attractive feature of being able to potentially combine variety of different data sets obtained by various methods into one set of experimental points placed in a convex hull of the network building blocks. These universality properties of the method require further studies and possibly a development of additional statistical methodology beyond the scope of our present work. In the current paper our main goal was to present a proof-of-concept example based on simulated data, with a purposefully straightforward but non-trivial model discrimination problem. For the example provided in this paper the method was seen to perform very well, with almost perfect discrimination against incorrect models even as the complexity of the model

# reactions	# cones	# non-degenerate cones	# building blocks	avg. running time	avg. success rate
m=5	5	5	6	4 sec	100%
m=6	15	11	15	30 sec	95%
m=7	35	30	133	6 min	94%
m=8	70	64	871	1h	97%
m=9	126	115	2397	8.5 h	96%

Table 2: Summary of average numerical results obtained from multiple experiments, each with $k = 10$ data points, $N = 2000$ rays in the Monte Carlo relative volume computation, and 2^{m-1} optimizations with random initial guess. The analysis was performed on a 2.8 Ghz Intel Core2Duo iMac machine.

selection problem increased.

Nonetheless, further studies and developments are needed to assess how well the method may perform on more challenging and realistic data sets. In particular, one of the aspects of the methodology which was not pursued here, and which could improve its computational scalability, is the utilization of techniques from computational algebra in order to increase the efficiency and further automate the proposed maximization algorithm.

Acknowledgements. The authors would like to thank Peter Huggins and Ruriko Yoshida for very helpful discussions, and additionally thank Peter Huggins for making available his Matlab script for volume calculations. The research was partially sponsored by the "Focused Research Group" grants NSF–DMS 0840695 (Rempala) and NSF–DMS 0553687 (Craciun) as well as by the NIH grant 1R01DE019243-01 (Rempala) and the NIH grant 1R01GM086881-01 (Craciun).

References

[1] M. Bansal, V. Belcastro, A. Ambesi-Impiombato and D. di Bernardo How to infer gene networks from expression profiles. *Molecular Systems Biology* 3:78 (2007). *www.molecularsystemsbiology.com DOI:10.1038/msb4100120* .

[2] G. Craciun, C. Pantea, Identifiability of chemical reaction networks, *Journal of Mathematical Chemistry* 44:1, 244-259, 2008.

[3] E.J. Crampin, S. Schnell, and P. E. McSharry, Mathematical and computational techniques to deduce complex biochemical reaction mechanisms, *Prog. Biophys. Mol. Biol.* 86 (2004) 177.

[4] P. Erdi and J. Toth, *Mathematical Models of Chemical Reactions: Theory and Applications of Deterministic and Stochastic Models*, (Princeton University Press, 1989)

[5] I.R Epstein and J.A. Pojman, *An Introduction to Nonlinear Chemical Dynamics: Oscillations, Waves, Patterns, and Chaos*, Oxford University Press, 2002.

[6] Ethier, S. N. and Kurtz, T. G. (1986). *Markov processes.* Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York.

[7] L. Fay and A. Balogh, Determination of reaction order and rate constants on the basis of the parameter estimation of differential equations, *Acta Chim. Acad. Sci. Hun.* 57:4 (1968) 391.

[8] D.M. Himmeblau, C.R. Jones and K.B. Bischoff, Determination of rate constants for complex kinetic models, *Ind. Eng. Chem. Fundam.* 6:4 (1967) 539.

[9] L.H. Hosten, A comparative study of short cut procedures for parameter estimation in differential equations, *Computers and Chemical Engineering* 3 (1979) 117.

[10] Huggins, P and Yoshida, R. (2008) First steps toward the geometry of cophylogeny. Manuscript, available at `oai:arXiv.org:0809.1908`.

[11] A. Karnaukhov, E. Karnaukhova and J. Williamson, Numerical Matrices Method for Nonlinear System Identification and Description of Dynamics of Biochemical Reaction Networks, *Biophys. J.* 92 (2007) 3459.

[12] G. Maria, A review of algorithms and trends in kinetic model identification for chemical and biochemical systems, *Chem. Biochem. Eng. Q.* 18:3 (2004) 195.

[13] A. Margolini and A. Califano Theory and Limitations of Genetic Networks Inference from Microarray Data *Annals N.Y. Acad Sci. 1115: 51–72 (2007). www.interscience.wiley.com DOI:10.1002/sim.3017.*

[14] L. Pachter, B. Sturmfels, *Algebraic Statistics for Computational Biology*, Cambridge University Press, 2005.

[15] Rempala, G. A., Ramos, K. S., and Kalbfleisch, T. (2006). A stochastic model of gene transcription: An application to L1 retrotransposition events. *J. Theoretical Biology*, 242(1):101–116.

[16] T. Richardson and P. Spirtes (2002). Ancestral graph Markov Models. *Annals of Statistics 30:962–1030*

[17] R.T. Rockafellar, *Convex Analysis*, Princeton, NJ, 1970.

[18] E. Rudakov, Differential methods of determination of rate constants of noncomplicated chemical reactions, *Kinetics and Catalysis* 1 (1960) 177.

[19] E. Rudakov, Determination of rate constants. Method of support function, *Kinetics and Catalysis* 11 (1970) 228.

[20] S. Schuster, C. Hilgetag, J.H. Woods and D.A. Fell, Reaction routes in biochemical reaction systems: algebraic properties, validated calculation procedure and example from nucleotide metabolism, *J. Math. Biol.* 45 (2002) 153.

[21] S. Vajda, P. Valko and A. Yermakova, A direct-indirect procedure for estimating kinetic parameters, *Computers and Chemical Engineering 10* (1986) 49.