Algebraic Statistical Model for Biochemical Network Inference

Gheorghe Craciun[1], Jaejik Kim[2], Casian Pantea[1], and Grzegorz A. Rempala[2⋆]

[1] Department of Mathematics and Department of Biomolecular Chemistry, University of Wisconsin-Madison, Madison, WI 53706 craciun@math.wisc.edu, pantea@math.wisc.edu
[2] Department of Biostatistics and Cancer Research Center, Georgia Health Sciences University, Augusta, GA 30912 jaekim@georgiahealth.edu, grempala@georgiahealth.edu

Abstract. We describe a statistical method for predicting most likely reactions in a biochemical reaction network from the longitudinal data on species concentrations. Such data is relatively easily available in biochemical laboratories, for instance, via the popular RT-PCR technology. Under the assumed kinetics of the law of mass action, we also propose the data-based procedures for (i) estimating the prediction errors and (ii) network dimension reduction. The algorithm in (ii) allows in particular for the application of the original algebraic inferential procedure described in [3] without the unnecessary restrictions on the dimension of the network stoichiometric space. Simulated examples of biochemical networks are analyzed in order to assess the proposed methods' performance.

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

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

1 Introduction

Modern biological research often involves collecting detailed longitudinal data on biochemical species concentration [5, 19] for the purpose of extracting information on the structure of a network of biochemical reactions. Typically, it is assumed that the identity of the chemical species present in the network is known, and the goal of the structure inference is to identify the species interactions [20] under some pre-imposed dynamics law. This problem is of particular interest in the context of molecular and systems biology under the *law of mass action* dynamics, and as such has received considerable attention in the literature [2, 9, 13, 14, 16, 24–26, 28]. It appears that in many cases one may infer identical mass-action models for distinct networks even with the experimental data on all species concentration being of arbitrarily high accuracy (i.e., with no measurement error) and arbitrary temporal resolution (i.e., with any number of time points). This lack of uniqueness of chemical networks is sometimes referred to as the *"fundamental dogma of chemical kinetics"* [5–7]. The necessary and sufficient conditions for two distinct reaction networks to give rise to the same *deterministic* mass action model are described in a recent paper [4] where the problem of identifiability of mass action reaction networks is treated in detail. The key point made in [4] is that, if we

⋆ To whom correspondence should be addressed

identify the reactions by their stoichiometric vectors (as discussed in the next section), it is possible for different sets of such vectors to span the same positive cones, or at least to span positive cones that have a non-empty intersection (see, e.g., Figure 1 in [4] for a simple example). In such cases it is not possible to uniquely identify the spanning reactions, and hence the corresponding network, from the experimental data via the deterministic model alone.

Further complicating the identifiability problem is the fact that the experimental measurements for the study of a specific reaction network or pathway are often collected under many different experimental conditions, which affects the values of reaction rate parameters. Moreover, the reactions of interest may not be "elementary reactions" for which the reaction rates parameters must be constant, but the so-called "overall reactions" consisting of multiple elementary reaction steps combined into a single one. The reaction rate parameters may therefore reflect the concentrations of biochemical species that have not been included explicitly in the model and are *not* constant, but rather depend on specific experimental conditions, such as concentrations of enzymes and other intermediate species. Consequently, the estimated values of reaction rate parameters in the same network obtained in several different experiments may be quite different numerically, and not readily center around specific values. However, under the identical set of network reactions, the regions of observed estimate values should agree with the regions of the stoichiometric space spanned by specific reaction cones. Given multiple sets of data arising from the same *stochastic* network, it should therefore be possible to identify the most likely sets of such cones and hence also the sets of corresponding reactions.

Based on such geometric considerations a likelihood-based method for reaction sets identification was recently proposed in [3]. The method uses the sparse likelihood parametrization of a *multinomial algebraic model* (see, e.g., Chapter 1 in [22]), and relies on mapping the estimated reaction parameters into an appropriate convex region in the span of the network's reaction vectors. This approach reduces a network identification problem to a statistical inference problem for parameters of a multinomial distribution, which may then be solved using classical likelihood methods assisted with some recent ideas on computational analysis of convex polytops (see, e.g., [15]). Despite its attractiveness in many respects, there were two main deficiencies of the original model presented in [3]. The first was the reliance on a (restrictive) assumption of full rank of the network's stoichiometric matrix and the second was the lack of explicit estimates of the inferential errors. In the proof-of-concept examples given in [3] such estimates were obtained via simulations from the known true models, but this approach is clearly not available in practical circumstances when the true networks are, in fact, unknown.

The purpose of the current paper is to enhance the usability of the original model of [3] for inferential purpose by addressing the above deficiencies. As we describe it below, the requirement of the full stoichiometric rank may be circumvented by applying a data-driven pre-processing step which reduces the dimension of the original stoichiometric space. Similarly, the data may be also used to inform the error estimates of the predictions, with a semi-parametric method utilizing properties of the stochastic biochemical network models under the law of mass action. The details of the proposed algorithms, along with examples, are provided in Sections 3 and 2, respectively. As seen below, we found it convenient to draw on the computational techniques borrowed from the theory of stochastic kinetics as it allowed us to better account for both the measurement error and rate parameters variability discussed above. Further discussion is deferred to Section 4 which also contains a summary of the paper's main points and offers some conclusions. The implementations of all the algorithms discussed is provided as part of the "Bioreactor" software suite and are currently available at *https://neyman.georgiahealth.edu/Bioreactor.html*.

2 Algebraic Multinomial Model

In this section we discuss in detail the algorithm for assessing the prediction variability in our biochemical network inferential procedure. We start by briefly reviewing some of the main elements of the framework introduced in [3].
Consider a reaction network model with d biochemical species $\mathbf{A} = \{A_1, \ldots, A_d\}$ and m possible reactions among them. The set $\mathbf{A}$ may be regarded as a basis of $\mathbb{R}^d$ (see, for instance, Feinberg's lecture notes [11]). Then the particular formal sums of species on either side of a reaction, called *complexes*, may also be viewed as vectors of $\mathbb{R}^d$, where the s-th coordinate of a complex $C \in \mathbb{R}^d$ is equal to the number of molecules of species A_s in C. The *reaction vector* of a reaction $C_1 \to C_2$ is the vector $C_2 - C_1 \in \mathbb{R}^d$ and the linear subspace of $\mathbb{R}^d$ spanned by all the reaction vectors corresponding to the reactions in a network is called the *stoichiometric subpace* of that network. As mentioned above, we consider a reaction network with m reactions. In what follows, the corresponding set of reaction vectors will be denoted by $\mathbf{R} = \{R_1, \ldots, R_m\}$. As described in [4], the analysis of the reaction networks may be decomposed into the analysis of sub-networks (or, equivalently, stoichiometric subspaces) having a single source complex, i.e., forming a cone in the species space $\mathbb{R}^d$. Consequently, throughout the paper we restrict our attention to such "conic" networks only.

2.1 Multinomial Likelihood

Let $\mathcal{R}_d$ denote the collection of all $\binom{m}{d}$ positive cones spanned by subsets of d vectors of $\mathbf{R}$. Denote by $cone(\mathbf{R})$ the positive cone generated by the reaction vectors in $\mathbf{R}$. Let S be the partition of $cone(\mathbf{R})$ 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 the *building blocks* or the *partition chambers*.
Let Δ_{m-1} be a probability simplex in $\mathbb{R}^m$ and let $\boldsymbol{\theta} \in \Delta_{m-1}$ be a vector of probabilities associated with the reactions that give rise to $\mathbf{R}$. We assume that these m reactions have the same source complex (i.e., form a conic network, see above). Define the polynomial map

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

where

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

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

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

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

[1] In general, it may be beneficial to consider various measures $vol(\cdot)$ that are absolutely continuous with respect to the usual Lebesgue measure. For instance, in our numerical examples in the next section, we define this measure via gamma densities.

that the map p is rational but, as explained in [3], the model may be re-parameterized into an equivalent one involving only the simpler, 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

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

In order to find the data points u_i we assume that the reaction network follows the stochastic law of mass actions (see e.g., [12]) whose stochastic trajectories fluctuate around the ODE system $G(\cdot)$ of the form

$$d\mathbf{A}_t^O/dt = G(\gamma, \mathbf{A}_t^O) \tag{4}$$

where $\mathbf{A}_t^O$ are the concentrations of $\mathbf{A}$ at time t and γ are unknown coefficients, estimable from the observed species concentration data. It is important to note that for the conic networks the ODE operator $G(\gamma, \mathbf{A}_t^O)$ depends on $\mathbf{A}_t^O$ only through the values of the source species (see, e.g., (11) below). As described in [3], we assume that such data consists of n sets of concentrations (trajectories), possibly corresponding to different γ parameters. The estimate of γ based on the k-th set of concentrations, denoted $\hat{\gamma}_k$, $(k = 1, \ldots, J$, where $J = \sum_{i=1}^{n} u_i)$ is a d-dimensional vector which maps the chemical reaction rates into the species space. Each $\hat{\gamma}_k$ corresponds therefore to a "data point" in the species space and the collection of all such points cumulated over the distinct chambers gives the values of u_i in (3). Here and elsewhere in this paper (albeit this does not have to be the case in general) the $\hat{\gamma}_k$'s are the least-squares estimates of the ODE coefficients in the deterministic law of mass action model (4), that is, the linear combinations of the (unknown) reaction rate constants. Given the values $\hat{\gamma}_k$, the most likely reaction vectors are inferred based on their *maximum likelihood estimators* or MLEs, that is, the values of $\hat{\boldsymbol{\theta}}$ that maximize the likelihood function. More precisely, the inference problem is to find

$$\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) \qquad \text{subject to} \quad \sum_{i=1}^{m} \theta_i = 1 \quad \text{and} \quad \theta_i \geq 0. \tag{5}$$

Once the above problem is solved, the reactions corresponding to the indices j for which

$$\hat{\theta}_j \approx 0 \tag{6}$$

may be *removed* from the collection of the reaction vectors $\mathbf{R}$. Formally, this requires testing of various statistical hypothesis of the form

$$H_0 : \theta_{i_1} = 0, \theta_{i_2} = 0, \ldots, \theta_{i_k} = 0,$$

indexed by k-tuples of integers $(i_1, \ldots, i_k)$, $k \leq m$. The first issue we consider in the current paper is how to performing a data-based test of the above hypothesis by inverting the joined confidence region for θ_i's. This requires a method for assessing the variability of $\hat{\boldsymbol{\theta}}$ for given J data points $\hat{\gamma}$.

2.2 Variability Assessment

We propose the following two-step procedure of analyzing the variability of the multivariate estimate $\hat{\boldsymbol{\theta}}$. In the first step estimate the variability of the data-estimated values $\hat{\gamma}_k$, treated as empirically obtained points in the species space and, in the second step, use them to estimate the variability of the MLE $\hat{\boldsymbol{\theta}}$. The estimates $\hat{\gamma}_k$, may be, for instance, obtained via a simple and robust *least squares* minimization, which turns out in this case to be asymptotically equivalent

to a more computationally expensive likelihood maximization (see, e.g., [8] Chapter 10). The difficulty lies in the fact that the distribution of $\hat{\gamma} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_d)^\top$ is hard to describe, due to both the longitudinal nature of the data and its often very complex pattern of stochastic fluctuations ([8] Chapter 10) around the ODE system (4).
In view of the complicated nature of the underlying probability laws, in order to obtain the distributions of the least squares estimates (LSEs) we apply the block-bootstrap method (cf. e.g., [17]) in the semiparametric model

$$\mathbf{A}_t = \mathbf{A}_t^O + \sigma(\mathbf{A}_{t-1})\boldsymbol{\epsilon}_t \quad t = 0, 1 \ldots, T. \tag{7}$$

Here $\mathbf{A}_t = (A_{1t}, \ldots, A_{dt})^\top$ is a vector of concentrations of d species $A_1, \ldots, A_d$ at the equidistant[2] times t and $\mathbf{A}_t^O = (A_{1t}^O, \ldots, A_{dt}^O)^\top$ is a vector of the corresponding solutions of the ODE system of d differential equations (4) with rate coefficients $\gamma = (\gamma_1, \ldots, \gamma_d)^\top$. Furthermore, $\sigma(\mathbf{A}_{t-1}) = Diag(\sigma_1(A_{1,t-1}), \ldots, \sigma_d(A_{d,t-1}))$ is a $d \times d$ diagonal matrix with unknown diagonal elements $\sigma(A_{it})$ $(i = 1, \ldots, d)$, and $\boldsymbol{\epsilon}_t = (\epsilon_{1t}, \ldots, \epsilon_{dt})^\top$ is a random vector with zero mean, the covariance matrix consisting of unit diagonal entries, and unknown off-diagonal entries. Note that $\mathbf{A}_0$ is a known initial vector and $\hat{\mathbf{A}}_0 = \mathbf{A}_0$. Since we use a set of estimates $\hat{\gamma}_k$ of the differential equation coefficients as inputs in our inferential procedure, the model (7) includes $\mathbf{A}_t^O$, the solution of the ODE (4), as the mean function of the species concentration trajectories. Note that due to the Markov property of the trajectories assumed under our stochastic chemical kinetics (cf. e.g., [3]), the diagonal elements $\sigma_i(A_{i,t-1}),\ i = 1, \ldots, d$, depend only on the species concentrations at the time-points $t-1$ and t.
Whereas the quantities $A_{it}^O,\ i = 1, \ldots, d$ may be estimated by solving the least squares problem and substituting $\hat{\gamma}_k$ $(k = 1 \ldots, n)$ into the differential equation trajectories, the non-explicit form of $\sigma_i(A_{i,t-1}),\ i = 1, \ldots, d$ requires a non-parametric approach, offered for instance, by the Nadaraya-Watson type estimator [21]. Since the Markov property implies first-order autoregressive processes, the Nadaraya-Watson AR(1) model may be used to estimate the variance functions of the error terms as follows:

$$\hat{\sigma}_{i,h_i}^2(a) = \frac{(\hat{p}_{h_i}(a))^{-1}}{T-1} \sum_{t=2}^{T} K_{h_i}(a - A_{i,t-1})(A_{it} - \hat{A}_{it}^O)^2,\ i = 1, \ldots, d, \tag{8}$$

where

$$\hat{p}_{h_i}(a) = \frac{1}{T-1} \sum_{t=2}^{T} K_{h_i}(a - A_{i,t-1}).$$

Here $K_{h_i}(\cdot) = h_i^{-1} K(\cdot/h_i)$, with a kernel smoothing function, $K(\cdot)$ and a pre-specified bandwidth value for the i-th trajectory, h_i. The nonparametric estimate of the variance function (8) was recently proposed in [10]. From the estimates $\hat{\mathbf{A}}_t^O$ and $\hat{\sigma}(\mathbf{A}_{t-1})$, we can obtain the vector of residual $\mathbf{e}_t = (e_{1t}, \ldots, e_{dt})^\top$ estimating the errors $\boldsymbol{\epsilon}_t,\ t = 1, \ldots, T$. While $\boldsymbol{\epsilon}_t$'s have constant variances, they are likely correlated. In order to account for these correlations, we consider the *block bootstrap*, or block resampling, from $\mathbf{e}_t$ (see, e.g., [17]). In our current setup the resampling procedure above is typically biased, owning it to the bias of the LSEs of the ODE coefficients, and thus the estimated residuals corresponding to $\boldsymbol{\epsilon}_t$ in (7) do not have zero mean. To alleviate this effect we perform a bias correction by replacing the estimated residuals vector $\mathbf{e}_t$ with its corrected version $\mathbf{e}_t^{bc} = \mathbf{e}_t - \bar{\mathbf{e}}_t$, where $\bar{\mathbf{e}}_t = \sum_{t=1}^{T} \mathbf{e}_t/T$. The proposed bootstrap procedure may be then summarized as follows.

Algorithm 1 (Bootstrap Confidence Regions)

[2] The extension to non-equidistant time grid is straightforward but, for simplicity, not pursued here.

Suppose J sets of longitudinal concentration measurements at T time points for d species are obtained, each from a different experiment (i.e., representing a different model trajectory). Then, we have J observation matrices $\boldsymbol{\Psi}_j = [\mathbf{A}_{j0}, \ldots, \mathbf{A}_{jT}]^\top$, $j = 1, \ldots, J$ with $(T+1)$ rows and d columns. The algorithm proceeds in two steps.

1. *For each $\boldsymbol{\Psi}_j$ $j = 1, \ldots, J$;*

- ***(a)*** *Based on the entries $\boldsymbol{\Psi}_j$, the least-squares estimate $\hat{\gamma}_j = (\hat{\gamma}_{1,j} \ldots, \hat{\gamma}_{d,j})$ of the vector of parameters $\gamma_j = (\gamma_{1,j} \ldots, \gamma_{d,j})$ is obtained. The estimate $(\hat{\gamma}_j)$ is used to obtain the "plug-in" species concentration estimates $\hat{\mathbf{A}}^O_{jt}$, $t = 0, \ldots, T$ from the ODE model* (4)*. Note that $\hat{\mathbf{A}}^O_{j0} = \mathbf{A}_{j0}$.*
- ***(b)*** *Using the Nadaraya-Watson estimator (8), the estimates of $\hat{\sigma}(\mathbf{A}_{j,t-1})$ are obtained, and the residuals $\mathbf{e}_{jt} = (e_{1jt}, \ldots, e_{djt})^\top$, $t = 0, \ldots, T$ are computed, where $e_{i0} = 0$ and $e_{it} = (A_{it} - \hat{A}^O_{it})/\hat{\sigma}_i(A_{i,t-1})$, $t = 1, \ldots, T$. From $\mathbf{e}_t$, the bias corrected residuals are obtained by taking $\mathbf{e}^0_t = \mathbf{e}_t - \bar{\mathbf{e}}_t$, $t = 1, \ldots, T$, where $\bar{\mathbf{e}}_t = \sum_{t=1}^T \mathbf{e}_t/T$.*
- ***(c)*** *The circular block-bootstrap sample $\mathbf{e}^*_{jt}$, $t = 1, \ldots, T$ from $\mathbf{e}^0_t$ is obtained. The sample $\mathbf{e}^*_{jt}$ consists of a set of re-arranged blocks of observations of fixed size b generated from the ordered and bias corrected residuals $\mathbf{e}^0_{jt}$, $t = 1, \ldots, T$.*
- ***(d)*** *The set of resamples $\mathbf{A}^*_{jt}$ is obtained by taking*

$$\mathbf{A}^*_{jt} = \hat{\mathbf{A}}^O_{jt} + \hat{\sigma}(\mathbf{A}_{j,t-1})\mathbf{e}^*_{jt}, \ t = 0, \ldots, T. \tag{9}$$

- ***(e)*** *The vector of estimates for the bootstrap trajectories $\hat{\gamma}^*_j$ is calculated using the least squares method.*
- ***(f)*** *The steps (c)–(e) are repeated B times for the j-th observed trajectory, resulting in the set of resamples $\hat{\gamma}^*_{jk}$, $k = 1, \ldots, B$.*

2. *For $k = 1, \ldots, B$;*

- ***(a)*** *Consider D_k, a matrix with J rows and d columns that consists of the bootstrap estimates $\hat{\gamma}^*_{jk}$. i.e., $D_k = [\hat{\gamma}^*_{1k}, \ldots, \hat{\gamma}^*_{Jk}]$.*
- ***(b)*** *From the data matrix D_k, u^*_i is obtained as "pseudo-data" and used to derive the corresponding MLE $\hat{\boldsymbol{\theta}}^*$, i.e. a resampling realization of an estimate of $\boldsymbol{\theta}$. This is done by solving the problem* (5) *with u^*_i in place of u_i.*

Finally, the resampling estimate of the distribution of $\hat{\boldsymbol{\theta}} \in \Delta_{m-1}$ is calculated based on the B values $\hat{\boldsymbol{\theta}}^*$.

In order to provide a good approximation to the bootstrap distribution, as measured e.g., by the jackknife-after-bootstrap (JAB) method (cf. [18]), the number B of bootstrap resamples needs to be sufficiently large. To perform this bootstrap procedure for the j-th set of trajectories, the bandwidth of i-th species h_{ij} for the variance function in (8) and the block size b need to be selected. These are important tuning parameters, as their values may profoundly affect the quality of the final estimates. In all examples below, we take the Gaussian kernel smoothing function $K(\cdot)$ with the block size $b = T/10$ and $h_{ij} = (4/3)^5 s^A_{ij}(T+1)^{-1/5}$ for all i and j, where s^A_{ij} is the standard deviation of the concentrations $A_{ij0}, \ldots, A_{ijT}$. These values are based on [27] and seem to work well in our numerical examples below. Whereas more elaborate selection methods are also available, their discussion is outside our current scope. We illustrate the performance of Algorithm 1 with the following two examples.

Example 1. Consider the following network of four biochemical species $A_0, \ldots, A_3$

$$A_0 \to A_1, \quad A_0 \to A_1 + A_2, A_0 \to A_2 + A_3, \text{ and }, A_0 \to 2A_3. \tag{10}$$

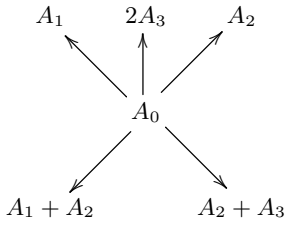
In the particular case of (10), the ODE system (4) becomes

$$dA_i/dt = \gamma_i A_0 \qquad i = 0, \ldots, 3. \tag{11}$$

where the parameters γ_i, $i = 0 \ldots 3$ are linear combinations of the true reaction rate constants. In this example our choice of theses rate constants yields $\gamma = (-1.1, 0.5, 0.8, 0.7)^\top$.
In order to asses the performance of our Algorithm 1, we simulated both the experimental and the validation data as the points on the species trajectories of the stochastic network (10) fluctuating around the deterministic system (11). All the data was obtained using the standard discrete event stochastic simulator [12]. In order to obtain the bootstrap estimates of the distribution of LSEs $\hat{\gamma}$ in (10), the first part of Algorithm 1 with $B = 1000$ resamples was applied to the single simulated set of T=20 species concentrations collected on the regular time grid. The bootstrap distribution of each component of $\hat{\gamma}$ was then compared with the empirical distribution based on the independently simulated set of $n = 1000$ mutually independent, full trajectories. The results are presented in Figure 1 as two sets of density plots. As may be readily seen, in terms of their shape, modality and variance, the bootstrap estimates (b) are seen to agree rather well with the empirical counterparts (a). The quantile-to-quantile plots (c) also indicate good agreement, except for slight residual bias of the block bootstrap, seen in the first and the last plot. This bias effect is, however, within the simulation study margin of error. □

In the next example we analyze the second part of Algorithm 1, concerned with the confidence bounds for $\hat{\boldsymbol{\theta}}$, the MLE solving the optimization problem (5).

Example 2. Consider the same network (10) of four biochemical species $A_0, \ldots, A_3$ but with an additional superfluous reaction $A_0 \to A_2$. The new network, containing both real and superfluous species interactions, is therefore



(12)

In order to assess the performance of the second part of Algorithm 1, we again used the data generated from the J independent stochastic trajectories of the true network (10). However, unlike in the previous example, now different sets the rate parameters were used for different trajectories, in order to simulate the experimental heterogeneity. This was achieved by drawing J sets of rates from the product of independent gamma random variables $\Gamma(\alpha = 1.5, \lambda = 1)$. For each of the J trajectories the corresponding estimates $\hat{\gamma}$ were calculated, based on independent sets of $T = 20$ longitudinal datapoints. As described in Section 2.1, each of these J values of $\hat{\gamma} \in \mathbb{R}^4$, treated as a point in the appropriate building block, contributed to the multinomial likelihood function (3).
Two scenarios, with $J = 10$ and $J = 20$ were considered, yielding the MLE values $\hat{\boldsymbol{\theta}} = (0.06,\ 0.43,\ 0.22,\ 0.29,\ 0.00)$ and $\hat{\boldsymbol{\theta}} = (0.25,\ 0.29,\ 0.10,\ 0.36,\ 0.00)$, respectively. Based on the data from these scenarios, Algorithm 1 with $B = 1000$ bootstrap resamples was used to identify the confidence regions for $\hat{\boldsymbol{\theta}}$. The results are summarized in Table 1 below where the marginal bootstrap means and marginal 95% confidence bounds for both scenarios are given.

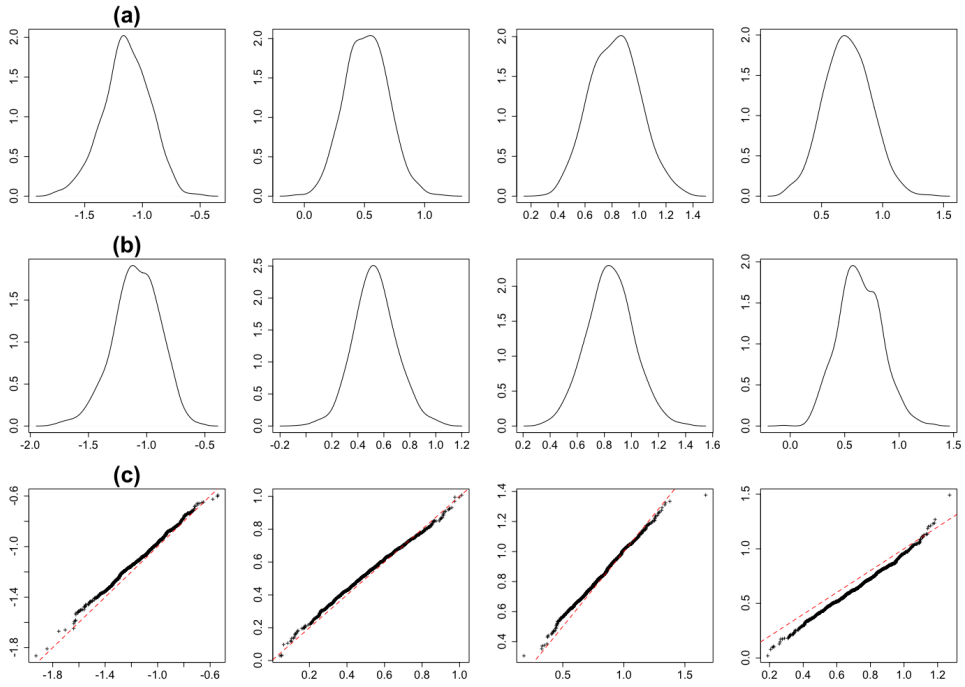


Fig. 1. (a) Marginal empirical distributions of LSE $\hat{\gamma} = (\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2, \hat{\gamma}_3)^\top$ obtained under the true model with $\gamma = (-1.1, 0.5, 0.8, 0.7)^\top$; (b) Matching marginal bootstrap distributions of $\hat{\gamma} = (\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2, \hat{\gamma}_3)^\top$ obtained with $B = 1000$ resamples; (c) Quantile plots for the matched pairs of (a) and (b) distributions compared against the equal quantile line (dashed).

Table 1. Bootstrap means and 95% confidence bounds (CB) for θ estimates based on $J = 10$ and $J = 20$ data points.

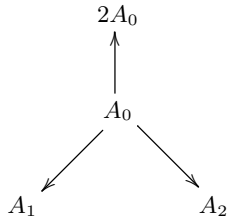
	$J = 10$		$J = 20$	
Reactions	Mean	95% CB	Mean	95% CB
$A_0 \to A_1$	0.09	[0.02,0.15]	0.21	[0.05,0.38]
$A_0 \to A_1 + A_2$	0.34	[0.20,0.52]	0.34	[0.18,0.55]
$A_0 \to A_2 + A_3$	0.23	[0.12,0.37]	0.10	[0.05,0.19]
$A_0 \to 2A_3$	0.34	[0.19,0.54]	0.34	[0.18,0.57]
$A_0 \to A_2$	0.00	[0.00,0.00]	0.00	[0.00,0.00]

As seen from the results in Table 1, the marginal confidence bounds for θ_i values corresponding to the first four (true) reactions all exclude zero value when $J = 10$ and $J = 20$, despite otherwise high uncertainty in the interval estimates due to relatively small sample size J. By the usual inversion of the confidence bound, the null hypothesis $H_{0i} : \theta_i = 0$ for $i = 1, \ldots, 4$ is therefore rejected at $\alpha = 5\%$ significance level. In contrast, the marginal 95% confidence bound for the fifth reaction probability θ_5 contains zero (in fact is concentrated at zero, at least up to two significant digits) and hence the null hypothesis $H_{05} : \theta_5 = 0$ is not rejected at $\alpha = 5\%$ level (nor, for that matter, at any level $\alpha \geq 1\%$). Consequently, the bootstrap analysis of the MLE confidence bounds suggests that, based on the observed data, the superfluous reaction should be removed from the network (12).

2.3 Dimension Reduction

The important assumption of the algebraic multinomial model (1)–(2), is that the span of the "true" reaction vectors is maximal, i.e., equals to d, the total number of species in the network. This is equivalent to the assumption of invertibility of the true network stoichiometric matrix and clearly holds for the network (10) considered in the examples above. If the stoichiometric matrix of the true network is not invertible, the multinomial likelihood method based on (5) may fail to meaningfully evaluate the reaction probabilities (in the sense of their significance assessment (6)). A simple example is as follows.

Example 3. Consider the network consisting of three possible reactions



(13)

with the true trajectory generating reactions

$$A_0 \to A_1 \text{ and } A_0 \to A_2. \tag{14}$$

Note that in the above network the cone spanned by the true reaction vectors ($[-1,\ 1,\ 0]$ and $[-1,\ 0,\ 1]$) has dimension two, while the space spanned by the species has dimension three. Figure 2 illustrated this clearly with the cloud of 500 data points simulated from

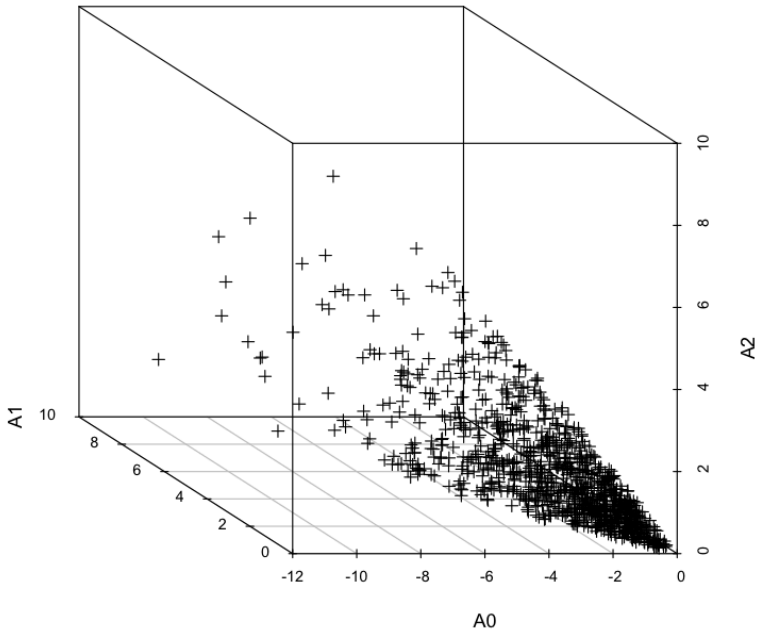


Fig. 2. The $\hat{\gamma}$ data points generated for the network $\{A_0 \to A_1, A_0 \to A_2\}$ obtained by the stochastic simulation procedure described in Example 1. Note that the points remain close to the positive cone spanned by the two reaction vectors. This cone has dimension two, while the space spanned by the species has dimension three.

the true reactions (14) via the method described in Example 2. As shown in Figure 2, the simulated data set of $\hat{\gamma}$ remains close to the plane $A_0 + A_1 + A_2 = 0$ determined by the stoichiometric vectors. Analyzing the likelihood estimates (5) based on $J = 500$ points we find the estimated probabilities as $\hat{\boldsymbol{\theta}} = (0.09,\ 0.16,\ 0.79)$. In this case the MLE, being forced into the three-dimensional species space, gives highest estimated probability value to the superfluous reaction $A_0 \to 2A_0$, which is clearly undesirable.

In order to extend the algebraic model (1)–(2) (and, consequently, also the bootstrap method of Algorithm 1) to networks with true stoichiometric space of reduced rank, we need a pre-processing step projecting the network model onto the appropriate species subspace. Whereas there are several possible ways of accomplishing this, we suggest below a particularly simple procedure based on the geometry of the $\hat{\gamma}$ points, reminiscent of the idea for the iterative angular principle component (see, e.g., [1]). The web-based implementation of the algorithm is available at *https://neyman.georgiahealth.edu/Bioreactor.html*

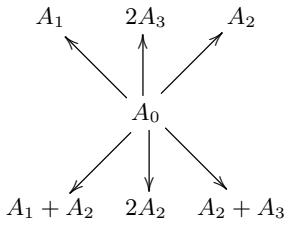
Algorithm 2 (Dimension reduction)

1. *Initialize the network $\mathcal{N}$ as equal to the full candidate reaction network, and denote by S the stoichiometric subspace of $\mathcal{N}$.*
2. *Let $k = \dim(S)$. Replace the data points $\hat{\gamma}$ by their orthogonal projections onto S, written in terms of the orthonormal basis of S.*

3. *Construct the list L of all $(k-1)$-dimensional subspaces generated by subsets of reaction vectors and compute the minimum α_{min} of all dihedral angles formed by disjoint pairs of subspaces in L.*
4. *For each subspace $s \in L$, compute the average of the angles between each data vector $\hat{\gamma}$ and its orthogonal projection on s.*
5. *Identify the subspace s^* which realizes the minimum average angle (say, δ) obtained in Step 4.*
6. *For a fixed threshold factor ϵ, if $\delta < \epsilon \cdot \alpha_{min}$ then discard the reactions whose reaction vectors do not lie in s^*. Update $\mathcal{N}$ to be the network formed by the remaining reactions, and S to be the stoichiometric subspace of $\mathcal{N}$, and go back to Step 2. If $\delta \geq \epsilon \cdot \alpha_{min}$ then stop.*

One can show that Algorithm 2 successfully determines the linear subspace that is spanned by the true reaction vectors, provided that the true values of the reaction rate constants do not have very different orders of magnitude, and that the measurement error for the $\hat{\gamma}$ values is not too high. An example of Algorithm 2 follows.

Example 4. Extend the network (12) in Example 2 to the following one



(15)

and suppose that the true reactions are

$$A_0 \to A_1 + A_2, \quad A_0 \to A_2 + A_3, \text{ and } A_0 \to 2A_3. \tag{16}$$

Two sets of $J = 10$ and $J = 20$ network trajectories for (15), are simulated from (16) as in Example 1. By design the data points are therefore confined to the proximity of a three-dimensional subspace of the four-dimensional real vector space with basis $\{A_0, A_1, A_2, A_3\}$. In both scenarios Algorithm 2 stops after two iterations as illustrated below by the output from the "Bioreactor" software (*https://neyman.georgiahealth.edu/Bioreactor.html*). The matrices below represent the reaction vectors in (15) as rows. In the initial matrix 'Rays' they are ordered clockwise, starting from the top-left in (15).

```
      −1 1 0 0
      −1 0 2 0                                        −1 0 2 0
      −1 0 0 2          −1 1 0 0          Possible:   −1 0 0 2
Rays :                  −1 0 1 0
      −1 0 1 0  Discarded:                            −1 0 1 1
      −1 0 1 1                                        −1 1 1 0
      −1 1 1 0
```

As seen, the dimension reduction algorithm has correctly concluded that the data points in the simulation should belong to the hyperplane $2A_0 + A_1 + A_2 + A_3 = 0$ and that the true reactions are among the four vectors included in this hyperplane, namely

$$A_0 \to 2A_2, \ A_0 \to 2A_3, \ A_0 \to A_2 + A_3, \ A_0 \to A_1 + A_2.$$

The dimension reduction algorithm produces a new set of data points that lie in the stoichiometric subspace S of the "reduced" network above and are written in coordinates corresponding to the orthonormal basis of S. For these new data points, the algebraic multinomial model (3) may be applied to estimate probabilities $\hat{\boldsymbol{\theta}}$, as before. Note, however, that $\hat{\boldsymbol{\theta}}$ estimated from the projected data points should now be interpreted as (conditional) probability in the reduced space.
The confidence bounds for the probabilities in the reduced space may be obtained by using Algorithm 1, but with the additional pre-processing via Algorithm 2 applied to each resample. The results for the current example, based on $B = 1000$ resamples are summarized in Table 2 below. As seen from the results summary the first three reactions in Table 2 have their

Table 2. Means and 95% confidence bounds (CB) for θ obtained using the dimension reduction method for $J = 10$ and 20 data points.

	$J = 10$		$J = 20$	
Reactions	Mean	95% CB	Mean	95% CB
$A_0 \to A_1 + A_2$	0.38	[0.26,0.54]	0.40	[0.27,0.57]
$A_0 \to A_2 + A_3$	0.18	[0.11,0.27]	0.18	[0.10,0.27]
$A_0 \to 2A_3$	0.44	[0.30,0.59]	0.42	[0.29,0.59]
$A_0 \to A_1$	0.00	[0.00,0.00]	0.00	[0.00,0.00]
$A_0 \to A_2$	0.00	[0.00,0.00]	0.00	[0.00,0.00]
$A_0 \to 2A_2$	0.00	[0.00,0.00]	0.00	[0.00,0.00]

respective 95% confidence bounds for θ separated from zero, which indicated that they should be considered significant in the network model (12). In contrast, the confidence bounds for the remaining reactions cannot be separated from zero (in fact, are concentrated at zero) up to two significant digits, suggesting that they are superfluous. This is consistent with the specified true network (16).

2.4 Measurement Error Effect

Throughout the paper we have assumed so far that the data on the trajectory is measured without errors. In order to assess the robustness of our Algorithm 2 against that assumption, we performed yet another simulation study, similar to the one in [3], in which the zero-mean Gaussian noise with varying standard deviation was added to a set of stochastic trajectories $\mathbf{A}_t$, $t = 1 \ldots, T$ corresponding to each species for the reaction network (15). Here we again used (15) as the hypothesized network and generated trajectory data from two different models of two and three reactions, respectively

$$2D: \ A_0 \to A_1 + A_2, \quad A_0 \to A_2 + A_3, \tag{17}$$

$$3D: \ A_0 \to A_1 + A_2, \quad A_0 \to A_2 + A_3, A_0 \to 2A_3. \tag{18}$$

From the trajectories with added noise, the LSE vector $(\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2, \hat{\gamma}_3)$ was obtained as in previous sections (i.e., $J = 20$ and $T = 20$) and Algorithm 2 was applied as before. The scenario was repeated in batches of 500 for each value of the standard deviation on the regular grid from 0 to .5. To reduce the computational overhead, we have assumed that the reactions with $\hat{\theta}_i \leq 0.005$ were identified as false. The result of this experiment is presented in Figure 3 in terms of the standard deviations (SDs) of the added Gaussian noise, plotted

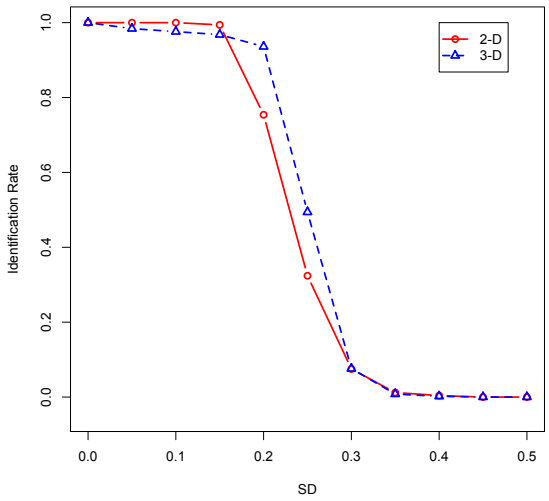


Fig. 3. The rate of recovery of the correct reactions (17) and (18) out of the network (15) as a function of the size of the Gaussian noise (in SD units) added to the trajectories.

versus the rates of correct reactions recovery (based on the batches of size 500). As seen from the plots, with SD up to 0.2, the maximum likelihood algorithm with dimension reduction was still able to recover the correct set of reactions at a remarkably high rate. However, the success rate drastically dropped down afterwards for both 2- and 3-dimensional case (17)–(18) and decreased rapidly to zero for SD above 0.3 or so (in fact, in our simulations the dimension reduction stopped to occur above that noise level). Since in our example the noise of SD 0.2 represents about 20% data value distortion, the dimensions reduction algorithm is seen as reasonably robust even in the presence of a relatively high measurement noise.

3 Summary and Discussion

The current paper expands on the likelihood-based algebraic statistical model, proposed recently in [3] for inferring biochemical reaction networks. The model allows us to infer the most likely network structure from the multiple sets of concentration data. The model provides a statistical solution to the (deterministic) non-identifiability problem caused by the fact that the different chemical reaction networks may give rise to exactly the same reaction rate equations. The attractiveness of the algebraic statistical method is in its ability to take advantage of the algebraic and geometric structure 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, for instance, Bayesian or probabilistic boolean networks (cf. [23]).

The present paper addresses some of the deficiencies of the original model described in [3] by providing an algorithm for reliable error bounds identification on the predicted network structure and by removing the restrictive assumption about the stoichiometric subspace of

the true network being fully-dimensional. As a result, a new non-parametric block-resampling procedure is proposed, allowing one to obtain confidence intervals for the reaction parameters and hence to assess the model predictions variability and precision. The approach proposed is versatile in the sense that it relies on relatively few assumptions about the reaction network, beyond the usually assumed stochastic law of mass action and the Markovian property. However, the drawback of the developed method, stemming from this very versatility, is its reliance on the empirical distribution of the data and, consequently, the computationally intensive resampling with proper parameter tuning and data pre-processing. The pre-processing dimension-reduction step is needed in order to alleviate the high computational cost as well as remove the somewhat restrictive assumption of a full rank of the stoichiometric matrix.
Overall, both in the examples provided as well as in other similar numerical experiments not reported here, we have found that the full algorithm presented in this paper (with the pre-processing step for dimension reduction) has performed very well, being able to recover the true network structure with high accuracy as long as the level of noise present in the concentration data was not too large (in our examples, less than 20%). Albeit these results are encouraging, further and more extensive studies, as well as more theoretical developments, are needed to further assess the statistical algebraic method's true practical applicability to large biochemical networks and to real experimental data. In future work, we intend to address these and other outstanding issues, such as linear independence of the reaction vectors and practical ways of performing the volume computations for large sets of reactions.

Availability. All algorithms described are implemented as part of the "Bioreactor" software suite available at *https://neyman.georgiahealth.edu/Bioreactor.html*.

Acknowledgments. This research was partially sponsored by the "Focused Research Group" grants NSF–DMS 0840695 (Rempala) and NSF–DMS 0553687 (Craciun) as well as by NIH grant R01DE019243 (Rempala) and NIH grant R01GM086881 (Craciun).

References

1. A. Altis, P. H. Nguyen, R. Hegger, and G. Stock Dihedral angle principal component analysis of molecular dynamics simulations *Journal of Chemical Physics 126, 244111* (2007) DOI:10.1063/1.2746330
2. 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)
3. G. Craciun, C. Pantea, and G. Rempala, Algebraic Methods for Inferring Biochemical Networks: a Maximum Likelihood Approach, *Computational Biology and Chemistry* 33(5):361-7 (2009) arXiv:0810.0561v2
4. G. Craciun and C. Pantea, Identifiability of chemical reaction networks, *Journal of Mathematical Chemistry* 44:1, 244-259, 2008.
5. E.J. Crampin, S. Schnell, and P.E. McSharry, Mathematical and computational techniques to deduce complex biochemical reaction mechanisms, *Progress in Biophysics and Molecular Biology* 86 (2004) 177.
6. P. Erdi and J. Toth, *Mathematical Models of Chemical Reactions: Theory and Applications of Deterministic and Stochastic Models*, (Princeton University Press, 1989)
7. I.R. Epstein and J.A. Pojman, *An Introduction to Nonlinear Chemical Dynamics: Oscillations, Waves, Patterns, and Chaos,* Oxford University Press, 2002.
8. S. Ethier and T. Kurtz *Markov Processes: Characterization and Convergence* (Wiley Series in Probability and Statistics) Wiley, New York 1986.

9. L. Fay and A. Balogh, Determination of reaction order and rate constants on the basis of the parameter estimation of differential equations, *Acta Chimica Academiae Scientarium Hungarica* 57:4, 391 (1968).
10. J. Franke, J-P. Kreiss and E. Mammen, Bootstrap of kernel smoothing in nonlinear time series. *Bernoulli* 8: 1–37 (2002).
11. M. Feinberg, Chemical reaction network structure and the stability of complex isothermal reactors II. Multiple steady states for networks of deficiency one, *Chemical Engineering Science*, 1:1–25, 43 (1988).
12. D.T. Gillespie Exact stochastic simulation of coupled chemical reactions, *Journal of Physical Chemistry* 81:2340–61, (1977).
13. D.M. Himmeblau, C.R. Jones, and K.B. Bischoff, Determination of rate constants for complex kinetic models, *Industrial and Engineering Chemistry Fundamentals* 6:4, 539 (1967).
14. L.H. Hosten, A comparative study of short cut procedures for parameter estimation in differential equations, *Computers and Chemical Engineering* 3, 117 (1979).
15. P. Huggins and R. Yoshida (2008). *First steps toward the geometry of cophylogeny.* Manuscript, available at `oai:arXiv.org:0809.1908`.
16. A. Karnaukhov, E. Karnaukhova and J. Williamson, Numerical Matrices Method for Nonlinear System Identification and Description of Dynamics of Biochemical Reaction Networks, *Biophysics Journal* 92, 3459 (2007).
17. S.N. Lahiri, *Resampling methods for dependent data,* New York: Springer-Verlag, 2003.
18. S.N. Lahiri, Consistency of the jackknife-after-bootstrap variance estimator for the bootstrap quantiles of a studentized statistic *Annals of Statistics* 33:5, 2475–2506 (2005).
19. G. Maria, A review of algorithms and trends in kinetic model identification for chemical and biochemical systems, *Chemical and Biochemical Engineering Quarterly* 18:3, 195 (2004).
20. A. Margolini and A. Califano, Theory and Limitations of Genetic Networks Inference from Microarray Data *Annals N.Y. Academy of Science* 1115: 51–72 (2007).
21. E.A. Nadaraya, On estimating regression, *Theory of Probability and its Applications*, 10: 189–190 (1964).
22. L. Pachter, B. Sturmfels, *Algebraic Statistics for Computational Biology,* Cambridge University Press, 2005.
23. T. Richardson and P. Spirtes Ancestral graph Markov Models. *Annals of Statistics* 30:962–1030 (2002).
24. E. Rudakov, Differential methods of determination of rate constants of noncomplicated chemical reactions, *Kinetics and Catalysis* 1, 177 (1960).
25. E. Rudakov, Determination of rate constants. Method of support function, *Kinetics and Catalysis* 11, 228 (1970).
26. 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, *Journal of Mathematical Biology* 45, 153 (2002).
27. B.W. Silverman, *Density Estimation,* Chapman and Hall, London, 1986.
28. S. Vajda, P. Valko and A. Yermakova, A direct-indirect procedure for estimating kinetic parameters, *Computers and Chemical Engineering* 10, 49 (1986).