MODEL REDUCTION IN BIOCHEMICAL REACTION KINETICS

B. Bamieh* **and L. Giarré†**

Abstract

We develop a method by which a large number of differential equations representing biochemical reaction kinetics may be represented by a smaller number of differential equations. The basis of our technique is a conjecture that the high dimension equations of biochemical kinetics, which involve reaction terms of specific forms, are actually implementing a low dimension system whose behavior requires right hand sides that can not be biochemically implemented. For systems that satisfy this conjecture, we develop a simple approximation scheme based on multilinear algebra that extracts the low dimensional system from simulations of the high dimension system. We demonstrate this technique on a standard 10 dimensional model of circadian oscillations and obtain a 3 dimensional sub-model that has the same rhythmic, birhythmic and chaotic behavior of the original model.

1 Introduction

The differential equations describing biochemical reaction networks often involve a large number of species (in the 10s or 100s). These equations are typically nonlinear and much of the understanding of these models currently comes from observing their behavior through numerical simulations. This is mainly due to the lack of analysis tools for high dimension nonlinear dynamical systems.

Biochemical kinetic equations can be written in the general form

$$\dot{x} \;=\; f(x), \tag{1}$$

where each of the state variables $x_i(t)$ represents the concentration of a reactant (species) at a given time t. An interesting special feature of kinetic equations is that the "rate function" f is typically (thought not always) polynomial (of several variables). Each component of f is a linear combination of multinomial terms of the form $\alpha x_{i_1} \dots x_{i_l}$, where α is the rate of the reaction between species $i_1, \dots, i_l$. It is often the case that each of these multinomial terms involve at most only two or three state variables since most constituent reactions are between only two or three species at a time. Thus, the function f in equation (1) is a polynomial of at most second or third degree. For the sake of motivation, in what follows we assume that the function f is of at most quadratic order.

One may think of a biochemical reaction network as being designed to produce a given behavior, say a system switching between equilibria or limit cycles depending on values of certain parameters. The observed species whose behavior is of interest are typically few, say two or three proteins. For such a small number of species, label them y_1, y_2 and y_3, it is often possible to "design" dynamics like

$$\dot{y} \;=\; P(y), \tag{2}$$

where P is a function that can be designed so that the system has a prescribed behavior. However, P will most likely not be a second order polynomial, but will involve many higher order terms and possibly irrational terms.

The above reasoning leads to the following conjecture. Biochemical reaction networks are trying to produce a behavior that can be captured by a low dimension system like (2), where P maybe a high order polynomial. However, the nature of biochemical kinetics is that it can only *realize* systems with a quadratic right hand side like (1), but with as many reactants as needed. It is thus plausible that the large number of extra species are there in order to realize the required behavior using quadratic biochemical kinetics.

If the conjecture is correct, then it is possible to take a high dimension biochemical kinetics equation like (1) and capture its behavior through a low dimensional system like (2). This comes at the expense of using a right

*B. Bamieh is with Department of Mechanical Engineering, University of California at Santa Barbara, USA bamieh@engineering.ucsb.edu

†L. Giarré is with DIEF, Università di Modena e Reggio Emilia, Italy laura.giarre@unimore.it

hand side P of a more complex form than f. This however is not a serious disadvantage for the following reason. The main tools for analyzing nonlinear dynamical systems are graphical. They involve analysis of phase portraits, from which equilibria, orbits and limit cycles can be identified. This is tractable for low dimension systems (dimensions of 2 or 3). Since one uses mainly graphical tools, the low order system (2) is thus more readily analyzed than (1) regardless of how complex the form of the function P is.

To give the previous discussion a more precise footing, we now present another motivation for our method that involves viewing common techniques like the Carleman linearization in reverse. It is well known [4; 7] that any nonlinear system of the form (2) can be *embedded* in an infinite dimensional linear system of the form

$$\dot{Y} \;=\; A\,Y, \tag{3}$$

where Y is an infinite dimensional vector with the original vector y as a subcomponent, and A is a linear operator. The system (2) is embedded in (3) in the sense that for any initial condition of (2) there is an initial condition of (3) where the trajectory of the z subcomponents of Z are identical to the trajectory of (2).

The Carleman linearization is a kind of global linearization scheme. There's a related procedure [5; 6] which we call *quadratization* in which a system like (2) can be embedded in a system of the form

$$\dot{Y} \;=\; \mathcal{Q}(Y), \tag{4}$$

where the function $\mathcal{Q}$ is a quadratic polynomial. When the original function P is polynomial (of any order), the quadratization (4) is finite dimensional.

The common feature of both the Carleman linearization and quadratization procedures above is that they take a lower dimension system with a complex right hand side P, and embed it in a higher dimensional system with a simpler right hand side (linear AY or quadratic $\mathcal{Q}(Y)$ in the case of Carleman linearization and quadratization respectively). The idea we pursue in this paper is to see whether such procedures (the quadratization in particular) can be reversed. In other words, given a high dimensional system with a quadratic right hand side like (4), one can ask whether it is the quadratization of a low dimensional system like (2) with a possibly complex right hand side P. The utility of this comes from the fact that if the embedded system (2) is of low enough dimension (say 2 or 3), then it can be analyzed using graphical tools, and its behavior understood without the need for extensive simulations.

In this paper, we do not directly address the mathematical problem posed in the previous paragraph. Instead, we use a function-fitting based technique in which important trajectories of the high dimension system are generated by simulation, and then a low dimension model is found that fits a portion of the data. This procedure involves two heuristics. The first is the selection of initial conditions for the original model that generate "important" or representative trajectories. The second heuristic is to choose a small number of states whose behavior is considered representative of the overall behavior of the original system. Once these choices are made, a function fit for the right hand side of (2) can be performed using the trajectory data. We show that by using tensor algebra that this function fit can be reduced to solving a large system of linear equations when the function P is a polynomial of several variables. This idea was preliminarly presented in [9]. Discovering governing equations from data by sparse identification has been published recently in [10], where the main focus stands into the identification of the underlying structure of a nonlinear dynamical system from data.

2 Model Reduction Approach

Let us consider a high order nonlinear system described by a system of Ordinary Differential Equations in the state variable $x = [x_1 \dots x_N]' \in \mathbb{R}^N$

$$\dot{x} = f(x). \tag{5}$$

We assume that we can identify, from physical a priori knowledge, a subset of the states whose behavior is to be studied. Let

$$y \;:=\; [y_1 \dots y_n]' \;=\; [x_{1_j} \dots x_{i_j}]', \qquad i_j \in [1 \dots N], \tag{6}$$

be such a subset of the states. The situation we are interested in is when n is small, i.e. 2, 3 or possibly 4.

The assumption we make is that the dynamics of each y_j can be written as

$$\dot{y}_j = f_{i_j}(x) = g_j(y), \tag{7}$$

for some function g_j which may be of higher order than f_{i_j}. This is the main content of the conjecture stated in the introduction; that the role of extra states is to enable the realization of the dynamics of the $y's$ using a function f which is only quadratic. Given this assumption we now outline a function fitting procedure that finds the parameters of the unknown polynomial functions $g_j(y)$. The function fitting procedure is most conveniently described using tensor algebra and Kronecker product notation. Given a vector x, define the following hierarchy of Kronecker products [4]

$$\begin{aligned} x^{(1)} &= x \\ x^{(2)} &= x \otimes x \\ x^{(3)} &= x \otimes x^{(2)} \\ &\vdots \end{aligned} \tag{8}$$

where if $x = [x_1\ x_2\ x_3]'$, then $x \otimes x = x^{(2)} = [x_1^2\ x_1x_2\ x_1x_3\ x_2x_1\ x_2^2\ x_2x_3\ x_3x_1\ x_3x_2\ x_3^2]'$, etc.

As is well known, polynomial functions can be represented as linear operators on tensor product spaces. If g is a polynomial function of degree m (defined as the highest degree of all its multinomial terms), then it can be written as the matrix vector product

$$g(y) \ = \ G\,Y,$$

where

$$Y \ := \ [1\ y'\ y^{(2)\prime} \dots y^{(m)\prime}]', \tag{9}$$

where G is a matrix whose entries correspond to the coefficients of the multinomial terms in g. We note that this is not the most compact representation of a polynomial function since the tensors $y^{(l)}$ have a redundancy due to the symmetry of multinomial terms. However, this "non-minimal" representation allows for the use of very simple notation.

Now the differential equations (7) can be written in vector form as

$$\dot{y}(t) \ = \ g(y(t)) \ = \ G\,Y(t). \tag{10}$$

Since the relationship between $\dot{y}$ and Y is linear, we can determine the entries of G (and consequently g) by solving systems of linear equations with data for Y formed from the data for y according to (9). The procedure can be summarized as follows. Given some representative trajectories of the original system (5), a selection of a subset of the states (6), a selection of the order m, and a set of time points $\{t_1, \dots, t_T\}$, collect the samples

$$\begin{aligned} \phi &:= \left[\, \dot{y}(t_1) \cdots \dot{y}(t_T) \,\right] \\ \Phi &:= \left[\, Y(t_1) \cdots Y(t_T) \,\right]. \end{aligned}$$

The condition that this data comes from the trajectories of a system like (10) is equivalent to the matrix equation

$$\phi \ = \ G\,\Phi.$$

Thus we determine the system parameters from the linear least square problem

$$\hat{G} \ = \ \arg\min \ \|\phi \ - \ G\,\Phi\|_2, \tag{11}$$

where $\|.\|_2$ is the Frobenius norm on matrices. To better appreciate the dimensions of this least squares problem, we write out the entries in more detail as

$$\hat{G} = \arg\min \ \left\| \left[\dot{y}(t_1) \cdots \dot{y}(t_T) \right] \ - \right.$$

$$\left. G \begin{bmatrix} 1 & & 1 \\ y(t_1) & & y(t_T) \\ \vdots & \cdots & \vdots \\ y^{(m)}(t_1) & & y^{(m)}(t_T) \end{bmatrix} \right\|_2$$

The matrix ϕ has dimensions $n \times T$, while Φ has dimensions $k \times T$, where

$$k \ := \ \sum_{i=0}^{m} n^i, \tag{12}$$

is the dimension of the vector Y. One critical size parameter in this fitting problem is m, the order of the right hand side g of (10). This determines the number of variables in the problem. The fit error will decrease monotonically with increasing m. The other size parameter is T, the number of trajectory samples taken. This determines the number of equations in the problem. In general, the best choice is to use a set of samples for which the trajectories explore a significant portion of the state space. Since solving very large linear least squares problems is routine with modern numerical techniques, it is possible to solve the above problem for a large range of orders m and number of samples T.

3 Example: The Circadian oscillations of the PER/TIM proteins

As an example of biological application of our method, we have considered hereafter the Drosophila circadian oscillations in the levels of two proteins PER and TIM as resulting from the negative feedback exerted by a PER/TIM complex on the expression of the PER and TIM genes which code for these two proteins. In [2] on the basis of some experimental observations, the authors have proposed a theoretical model for the circadian oscillations of the PER and TIM proteins in Drosophila. They have observed the occurrence of chaos and birhythmicity by means of bifurcation diagrams and locate the different domains of complex oscillatory behavior in parameter space. This model has been largely used in literature to study its complex behavior (see [3], [8]).

The model consists of $N = 10$ state-equations corresponding to the pathways scheme reported in [3] of the model for circadian oscillations in Drosophila involving negative regulation of gene expression by a complex between PER and TIM. We will show that the proposed technique yields for this model very good agreement with a 3rd or 2nd order sub-model. The mentioned model with all the values used for the parameters is reported in the Appendix. From their analytical studies and from biological insight, it is clear that among the 10 species, only $n = 3$ are important, the state variables x_1, x_5 and x_{10}, corresponding to the protein PER, TIM and the complex PER/TIM. In [2], and in [8] it has been shown that for $v_{dt} = 2, v_{mt} = 0.99$ limit cycles can be predicted and that for $v_{dt} = 4.8$

and $v_{mt} = 0.28$ they showed chaos. Hereafter we consider the two setting of parameters. We refer to the first one as **case 1** and to the second one as **case 2**. For the **case 1** we consider two reduced models, one with two states (corresponding to the protein PER and TIM ($n = 2$) (**case 1.a** For the second case, in order to get the chaotic behavior, we consider the three states reduced model ($m = 3$). When $n = 3$, K, the number of term in the regressors becomes, for $m = 1 : 8$, $k(m) = 3, 12, 39, 120, 363, 1092, 3279, 9840$. When $n = 2$, K, the number of term in the regressors becomes, foe $m = 1 : 8$, $K(m) = 2, 6, 14, 30, 62, 126, 254, 510$. Running the algorithm, we need a trade off between the achievable approximation level ϵ and the grade of complexity of the approximate systems, let us choose an order $m = 5$. For $m = 1 : 5$ in the **case 1.a**, the obtained approximation error $E(m) = \|\phi - G\phi\|$ is $E(m) = 9.14,\ 2.09,\ 0.60,\ 0.38,\ 0.28$.

Finally, in the **case 2** we get $E(m) = 12.01,\ 2.55,\ 1.58,\ 1.16,\ 1.03$.

Moreover, let us note that due to the redundancy in the expression of Φ, the problem is rank deficient. For example, let us consider the case of three states $[y_1\ y_2\ y_3]'$, then for $m = 2$, only 9 terms are different among the $k = 13$ terms of Φ, and the corresponding matrix $\Phi'\Phi$ of dimension $k + 1 \times k + 1$ has rank 10. For $n = 3$ and $m = 1, \ldots, 5$ the effective rank is $4, 10, 20, 35, 56$. However, since we are solving a least squares problem, the rank deficiency does not fundamentally complicate the problem.

For $m = 5$, we report a portion of the data (the first 300 samples) for $i = 1, n$ and the corresponding fitting system, in the **case 1.a**, Fig. 1 and in the **case 2** in Fig. 2.

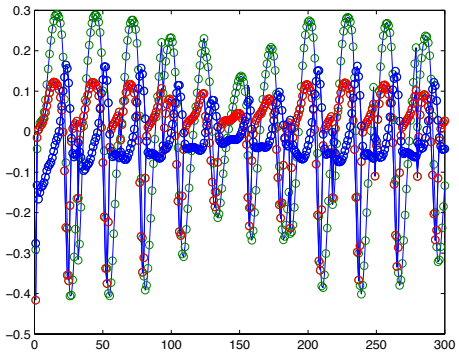

Figure 1. **case 1.b** Sample trajectories of the original system (solid lines) and the reduced 3'rd order system (dots)

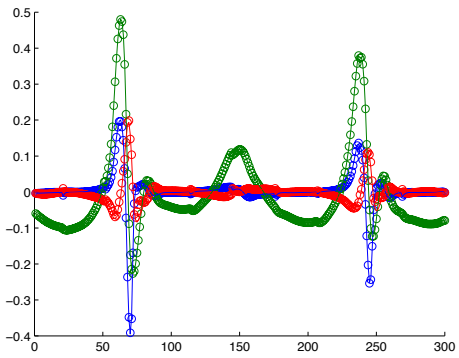

Figure 2. **case 2** Sample trajectories of the original system (solid lines) and the reduced 3'rd order system (dots)

References

Van De Wouw, H. Nijmeijer and D. Van Campen. A Volterra Series Approach to the Approximation of Stochastic Nonlinear Dynamics. *Nonlinear Dynamics*, 27:397-409, 2002.

J.C. Leloup and A. Goldbeter, Chaos and Birhythmicity in a Model for Circadian Oscillations of the PER and TIM Proteins in Drosophila, *Journal of Theor. Biol.*,198 (445-459) 1999.

J. C. Leloup and A. Goldbeter. A model for circadian rhythms in Drosophila incorporating the formation of a complex between the PER and TIM proteins. *Journ. Biol. Rhythms*, 2: 69-76, 1998

W. Rugh, *Nonlinear Systems Theory*, Originally published by The Johns Hopkins University Press,(ISBN O-8018-2549-0). 1981, Web version, 2002.

R. Stark1 and L. del Re. Low order representation of nonlinear systems. *Proc. of the 42nd IEEE Conference on Decision and Control* Maui, Hawaii USA, December 2003.

R. Starkl and L. del Re. Finite order representations of infinite dimensional bilinear models of nonlinear systems. *Proc. of the ACC*, 131-136, 2003.

W. Steeb. A note on Carleman Linearization. *Phisyscs Letters A*, 140, 6:336-338,1989.

Kunichika Tsumotoa, Tetsuya Yoshinagab, Hitoshi Iidac, Hiroshi Kawakamid, Kazuyuki Aiharae. Bifurcations in a mathematical model for circadian oscillations of clock genes. *Journal of Theoretical Biology*, 239 (2006) 101-122

B. Bamieh and L. Giarré. On Discovering Low Order Models in Biochemical Reaction Kinetics. *American Control Conference*, 2007.

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *PNAS* 2016 113 (15) 3932-3937; published ahead of print March 28, 2016, doi:10.1073/pnas.1517384113.