Proceedings of the 17th World Congress
The International Federation of Automatic Control
Seoul, Korea, July 6-11, 2008

Identification of Genetic Regulatory Networks: A Stochastic Hybrid Approach [⋆]

Eugenio Cinquemani [1], Andreas Milias-Argeitis and John Lygeros

Institut für Automatik, Eidgenössische Technische Hochschule Zürich, 8092 Switzerland.

Abstract:
Genetic regulatory networks are families of biochemically interacting genes that regulate most functions of a living cell via the synthesis of proteins and other essential molecules. In this paper we introduce a piecewise deterministic model of genetic network and devise a systematic procedure for the identification of the model parameters from experimental observations of the protein concentration dynamics. Numerical results on simulated data are presented to show the effectiveness of our method.

Keywords: Biological systems, Jump Markov systems, Parameter estimation, Uncertainty descriptions, Complex systems, Application of nonlinear analysis and design

1. INTRODUCTION

Genetic regulatory networks govern the synthesis of proteins and other essential molecules in the living cell, and are thus responsible for fundamental cell functions such as metabolism, development and replication. Thorough understanding of genetic networks is fundamental in that it determines our ability to interact with the basic biological mechanisms and to reproduce them.

Different approaches to genetic network modelling have been proposed in the literature and are conventionally classified into models with purely continuous dynamics and discrete event models de Jong (2002). However, it appears that certain systems are more naturally described by hybrid models that explicitly account for both continuous and discrete phenomena. This is witnessed by the number of researchers (Alur et al. (2001); de Jong et al. (2003); Drulhe et al. (2006); Batt et al. (2005); Ghosh and Tomlin (2004), among others) who recently applied hybrid systems tools in this context. In addition, the fundamental role of uncertainty in gene expression is being recognized, see for instance the work by Kaern et al. (2005); Vilar et al. (2002); McAdams and Arkin (1997, 1999).

Most recently, a number of researchers — Cinquemani et al. (2007); Drulhe et al. (2006); Perkins et al. (2004); Fujarewicz et al. (2005); Dunlop et al. (2007) — started to address the problem of learning genetic network models from data. The problem may be seen as a loop of three steps: 1) description of the interactions; 2) parameter identification; 3) validation. First step defines the structure of the network and typically provides a parametric model of gene expression dynamics. In step two, unknown parameters are estimated on the basis of experimental observations of gene expression. Finally, validation must evaluate the relevance of the resulting model to the observed system behavior, possibly providing hints on how to refine the network structure defined in step 1. In all of these steps, full exploitation of the experimental data is fundamental but very challenging. Gene expression levels are often observed at sparse, perhaps irregularly spaced observation times. In addition, different profiles are observed in several experiments of the same kind. This provides a large amount of information that requires careful processing.

The aim of this paper is to address parameter identification for genetic regulatory networks in a stochastic hybrid modelling framework. We introduce a class of piecewise deterministic models where protein concentrations follow first-order kinetics with synthesis rates that depend on the random activation of gene expression. In turn, gene expression follows the laws of a finite state Markov chain whose transition rates depend on the current protein concentration levels. For a given network of interactions, we consider the problem of estimating the unknown parameters of the model from protein concentration profiles. That is, we assume that step 1 above has been completed within our modelling framework and address step 2. We take advantage of the structure of the model to devise an algorithm that allows separate estimation of the parameters pertaining to different dynamical equations from convenient subsets of the observation data, with clear benefits in terms of computational complexity. In this procedure, the availability of multiple observations from independent experiments is explicitly taken into account.

In Section 2 we shall briefly discuss genetic interaction networks from a dynamic modelling perspective. A general stochastic hybrid model is introduced in Section 3. In Section 4 we state the parameter identification problem and derive the identification algorithm. The performance of the method is evaluated numerically on two simple case studies in Section 5. Concluding remarks and perspectives of our work are reported in Section 6.

[⋆] This research has been supported by the European Commission under project HYGEIA (NEST-4995).

[1] Corresponding author, e-mail: cinquemani@control.ee.ethz.ch.

978-1-1234-7890-2/08/$20.00 © 2008 IFAC

10.3182/20080706-5-KR-1001.1767

2. GENETIC NETWORKS

A genetic regulatory network consists of n interacting genes. A gene is a portion of the DNA that encodes one protein. Expression of a gene is activated (inhibited) by a certain activator (inhibitor) complex binding to a specific DNA motif. When a gene is expressed, several molecules of the encoded protein are synthesized through a process that includes several steps (e.g. transcription into mRNA and translation). On the other hand, the formation of the activation (inhibition) complex depends on the concentration of certain proteins of the network. Gene activation is a discrete event that involves the interaction of few molecules. Therefore it can be seen as a random event whose probability depends on the concentration of the proteins needed for the formation of the activation (inhibition) complex. On the other hand, the synthesis of several new molecules of a protein induces a modification in the protein concentration that can be described by simple synthesis/degradation laws. Note that the steps between the expression of one gene and the synthesis of the corresponding protein are implicitly lumped together. For the purpose of identification, this is not critical provided the experimental data consist of, or can be easily related to, protein concentration levels.

3. STOCHASTIC HYBRID MODEL

For convenience, we shall consider discrete-time dynamics. An equivalent continuous-time model may be set up along the lines of Kouretas et al. (2006). Let $x = [x_1, \dots x_n]^T \in \mathbb{R}_+^n$ be the vector of protein concentrations. For $i = 1, \dots, n$, the dynamics of concentration x_i is affine and described by a first-order linear equation: for $t \in \mathbb{N}$,

$$x_i(t+1) = a_i x_i(t) + f_i(t) + \bar{b}_i, \tag{1}$$

where $a_i \in (0, 1]$ is the degradation rate and $\bar{b}_i \geq 0$ is a fixed synthesis rate. Process $f_i(t) \geq 0$ is a variable synthesis rate that is expressed as a finite weighted combination of binary random processes:

$$f_i(t) = \sum_j b_{i,j} \prod_k u_{i,j,k}(t). \tag{2}$$

In turn, each process $u_{i,j,k}$ follows the laws of a first-order Markov chain whose transition probabilities depend on x_k:

$$\mathbb{P}[u_{i,j,k}(t+1) = 1 | u_{i,j,k}(t) = 0] = p_{i,j,k}\big(x_k(t)\big),$$
$$\mathbb{P}[u_{i,j,k}(t+1) = 0 | u_{i,j,k}(t) = 1] = q_{i,j,k}\big(x_k(t)\big).$$

It is assumed that, for $(i, j, k) \neq (i', j', k')$ and all t, $u_{i,j,k}(t)$ and $u_{i',j',k'}(t)$ are conditionally independent given $x_k(t)$ and $x_{k'}(t)$. Typically, not all components of $x(t)$ affect the laws of (2). We shall denote by $\ell(i) = [\ell_1, \dots, \ell_{l(i)}]^T \in \{1, \dots, n\}^{l(i)}$ the vector (with $l(i) \leq n$ distinct entries) containing the indexes of states that affect state i through (2). This is precisely the set of values taken on by k in (2). Thus, the laws of (2) are determined by $x_{\ell(i)} \triangleq [x_{\ell_1}, \dots, x_{\ell_{l(i)}}]^T$. We consider sigmoidal transition probabilities; that is, functions p and q take either of the following forms:

$$s^+(x_k; \eta, d) = \frac{x_k^d}{\eta^d + x_k^d}, \qquad s^-(x_k; \eta, d) = \frac{\eta^d}{\eta^d + x_k^d}.$$

Constant η determines the point x_k where the value of the sigmoid is 0.5 (hence we shall also call it *threshold value*), while the exponent d determines the steepness of the sigmoid. In general, constants η and d depend on the particular process $u_{i,j,k}$. Therefore, whenever essential we shall use the extensive notation $\eta_{i,j,k}$ and $d_{i,j,k}$. The case where $q_{i,j,k} = 1 - p_{i,j,k}$, for all i, j and k, appears to be relevant to the biological context, see Hu et al. (2004), and turns the Markov chain $u_{i,j,k}(t)$ into an independent process, cf. Cinquemani et al. (2007). From now on, we shall stick to this case. Extensions of our methods to the general case are rather straightforward and will not be discussed here. This model can be seen as a stochastic generalization of the piecewise affine model considered in Drulhe et al. (2006), and relates to the nonlinear model with sigmoidal regulation functions reviewed in de Jong (2002) when the expected evolution of the state is considered.

4. PARAMETER IDENTIFICATION

We assume that noisy observations of state x are taken at times $\tau \in \mathbb{N}_N$,where $\mathbb{N}_N \triangleq \{N, 2N, 3N, \dots\}$ (i.e. every N time steps) : for $i = 1, \dots, n$,

$$y_i(\tau) = x_i(\tau) + e_i(\tau), \tag{3}$$

where measurement error e_i is an i.i.d. process with mean zero and variance σ_i^2. Multiple statistically independent experiments will be considered. Therefore, for $m = 1, 2, \dots, M$ we shall write

$$y_i^m(\tau) = x_i^m(\tau) + e_i^m(\tau) \tag{4}$$

to denote the m-th of M experimental outcomes. As will be clarified later, for identification purposes, the assumption that all elements x_i of x are observed simultaneously in every experiment can be relaxed.

4.1 Problem statement

Suppose that the order n and the structure of the model (i.e. the specific form of (2), for $i = 1, \dots, n$) are given. Suppose in addition that parameters a_i, $\bar{b}_i$ and $b_{i,j}$ are known. We consider the following problem.

Problem 1. Given data from multiple experiments (4), estimate the unknown parameters $\theta_{i,j,k} = (\eta_{i,j,k}, d_{i,j,k})$.

That is, we assume that the interaction paths as well as the protein synthesis and degradation rate constants are known, and wish to learn the probability functions that govern the activation of gene expression. In fact, the identification method that we shall introduce can be extended to the problem of estimating all model parameters, including rate constants. Due to space limitations, though, this problem will not be addressed here.

4.2 Local approximate decoupling

Our aim is to devise an identification strategy that exploits the structure of the system to reduce the problem complexity. In particular, we wish to split identification into subproblems, each relying on a subset of the observation data and addressing estimation of a subset of the unknown coefficients. The starting point is the "quasi-diagonal" structure of the system:

$$\begin{bmatrix} x_1(t+1) \\ \vdots \\ x_n(t+1) \end{bmatrix} = \begin{bmatrix} a_1 & & 0 \\ & \ddots & \\ 0 & & a_n \end{bmatrix} \begin{bmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{bmatrix} + \begin{bmatrix} \bar{b}_1 \\ \vdots \\ \bar{b}_n \end{bmatrix} + \begin{bmatrix} f_1(t) \\ \vdots \\ f_n(t) \end{bmatrix}$$

where coupling between the n scalar equations is introduced indirectly by the rightmost term through the state-dependent laws of the $u_{i,j,k}$. Consider a point $\bar{x}$ in the state space. For any fixed i, let us define a stationary process

$$\bar{f}_i(t) = \sum_j b_{i,j} \prod_k \bar{u}_{i,j,k}(t), \tag{5}$$

with the same structure of (2), but governed by the time-invariant transition probabilities:

$$\mathbb{P}[\bar{u}_{i,j,k}(t+1) = 1 | \bar{u}_{i,j,k}(t) = 0] = p_{i,j,k}\big(\bar{x}_k\big),$$
$$\mathbb{P}[\bar{u}_{i,j,k}(t+1) = 0 | \bar{u}_{i,j,k}(t) = 1] = q_{i,j,k}\big(\bar{x}_k\big)$$

(with $q_{i,j,k} = 1 - p_{i,j,k}$). Now consider a time interval $\mathscr{T}$ such that

$$x_{\ell(i)}(t) \simeq \bar{x}_{\ell(i)}, \quad \forall t \in \mathscr{T} \tag{6}$$

(recall that $\ell(i)$ encodes the entries of x that affect f_i). Then we make the following approximation.

Approximation 1. Over the time interval $\mathscr{T}$,

$$f_i(t) \simeq \bar{f}_i(t) \tag{7}$$

in the sense of probability distribution.

Under (7), the dynamics of x_i are decoupled from the remaining components of the state vector, and are determined by the (entries $\ell(i)$ of the) approximation point $\bar{x}$. We shall rely on this local approximation to treat every equation $i = 1, \ldots, n$ separately.

Proposition 1. Under Approximation 1, for $t \in \mathscr{T}$, the following recursion holds:

$$\mathbb{E}[x_i(t+1)] = a_i \mathbb{E}[x_i(t)] + \bar{b}_i + \sum_j b_{i,j} \prod_k p_{i,j,k}(\bar{x}_k). \tag{8}$$

4.3 Estimation algorithm

Fix i. In light of the previous discussion, the idea is to estimate parameters $\theta_i \triangleq \{\theta_{i,j,k},\ \forall j,k\}$ by matching local statistics of x_i around $\bar{x}$ to empirical local statistics drawn from data collected "near" $\bar{x}$ (i.e. such that $y_{\ell(i)} \simeq \bar{x}_{\ell(i)}$). In fact, the behavior of the system at several different locations of the state space will be considered simultaneously. For $N_L = L \cdot N$, with $L \in \mathbb{N}$, and $\tau \in \mathbb{N}_N$, consider the (L-steps) expected variation of x_i:

$$\mathbb{E}[x_i(\tau + N_L) - a_i^{N_L} x_i(\tau)]/G_{N_L}(a_i), \tag{9}$$

where $G_{N_L}(a_i) = (1 - a_i^{N_L})/(1 - a_i)$.

Proposition 2. Let $\mathscr{T} = [\tau, \tau + N_L]$. Under Approximation 1, the expected variation (9) is given by the formula:

$$v_i(\bar{x}; \theta_i) \triangleq \bar{b}_i + \sum_j b_{i,j} \prod_k s(\bar{x}_k; \theta_{i,j,k}), \tag{10}$$

where $s = s^+$ or $s = s^-$ depending on i, j, k.

Proposition 2 shows us how to compute (9) as an explicit function of the unknown parameters θ_i in the case where $x_{\ell(i)}(t)$ is close to $\bar{x}_{\ell(i)}$ over a time interval spanning L subsequent observation times. An empirical counterpart of the expected variation of x_i about $\bar{x}$ may be computed from the data as follows. For a given tolerance vector $\delta \in \mathbb{R}_+^n$, consider a hyperrectangular neighborhood of $\bar{x}$:

$$\mathscr{X}_i(\bar{x}, \delta) \triangleq \{x \in \mathbb{R}_+^n :\ |x_{\ell(i)} - \bar{x}_{\ell(i)}| \leq \delta_{\ell(i)}\},$$

where $\delta_{\ell(i)} = [\delta_{\ell_1}, \ldots \delta_{\ell_{l(i)}}]^T$ and the inequality is interpreted componentwise. Let $\mathscr{M}_i(\bar{x}, \delta) \subset \{1, \ldots, M\}$ be the set of (indexes of) the observed trajectories such that $\exists \tau \in \mathbb{N}_N, \exists \lambda \geq 1$ for which

$$\{y^m(\tau), y^m(\tau + N), \ldots, y^m(\tau + \lambda N)\} \subset \mathscr{X}_i(\bar{x}, \delta).$$

For every $m \in \mathscr{M}_i(\bar{x}, \delta)$ let

$$L_m = \max_\tau \max \big\{\lambda \mid \{y^m(\tau), y^m(\tau + N), \ldots \\ \ldots, y^m(\tau + \lambda N)\} \subset \mathscr{X}_i(\bar{x}, \delta)\big\} \tag{11}$$

and let the maximum be attained at $\tau = \tau_m$. That is, τ_m and L_m together define the maximal piece of the m-th trajectory that lies in $\mathscr{X}_i(\bar{x}, \delta)$, provided $\lambda \geq 1$. Finally, let $M_i = \mathrm{card}(\mathscr{M}_i)$. Then, empirical variations of x_i about $\bar{x}$ are computed by the formula

$$\hat{v}_i(\bar{x}) \triangleq \frac{1}{M_i} \sum_{m \in \mathscr{M}_i} \frac{y_i^m(\tau_m + N_{L_m}) - a_i^{N_{L_m}} y_i^m(\tau_m)}{G_{N_{L_m}}(a_i)} \tag{12}$$

(compare this to (9) and observe that $\mathbb{E}[y_i] = \mathbb{E}[x_i]$). We are now ready to state our identification procedure. The algorithm below applies separately to all $i = 1, \ldots, n$.

Algorithm 1. (Local Approximate Decoupling).

```
• choose a positive integer $H$;
• choose points in the state space $\bar{x}^{(h)}$ and tolerance
  vectors $\delta^{(h)} \in \mathbb{R}_+^n$, with $h = 1, \ldots, H$;
• for $h = 1, \ldots, H$ do
    · compute set $\mathscr{M}_i(\bar{x}^{(h)}, \delta^{(h)})$ and the values of $\tau_m$
      and $L_m$, $\forall m \in \mathscr{M}_i(\bar{x}^{(h)}, \delta^{(h)})$;
    · compute $\hat{v}_i(\bar{x}^{(h)})$ as in (12);
• solve
      $\hat{\theta}_i = \arg\min_{\theta_i} \sum_{h=1}^{H} [\hat{v}_i(\bar{x}^{(h)}) - v_i(\bar{x}^{(h)}; \theta_i)]^2$,   (13)
  where the $v_i(\bar{x}^{(h)}; \theta_i)$ are evaluated by way of (10).
```

In practice, the empirical estimates $\hat{v}_i(\bar{x}^{(h)})$ at several points in the state space are regarded as noisy measurements of the underlying function $v_i(\bar{x}^{(h)}; \theta_i^*)$, where θ_i^* denotes the putative true values of the parameters. This leads to turning parameter estimation into the nonlinear regression expressed by (13). The choice of the approximation points $\bar{x}^{(h)}$ is fundamental, and should be driven by the structure of the system (which is assumed to be known). In particular, a sufficient number of points H should be considered to guarantee that optimization (13) is well defined.

The accuracy of the procedure depends on several factors. Noise variance σ_i^2 determines the uncertainty of estimates $\hat{v}_i$ and the probability that data sequences lying in the $\mathscr{X}_i$ correspond to state sequences within the same set. In principle, larger values of L reduce the variance of $\hat{v}_i$. Larger values of M generally increase the amount of data available per approximation point, i.e. the size of the $\mathscr{M}_i$, which in turn leads to a reduction of the uncertainty of $\hat{v}_i$. Finally, the validity of approximation (7) depends on the steepness on the sigmoidal transition probabilities: the steeper the functions, the smaller the regions of the state space where local approximation holds with given accuracy. Note that this factor cannot be quantified ahead of identification, in that steepness is determined by the unknown parameters themselves. A rigorous theoretical analysis of the performance of the method is currently being developed.

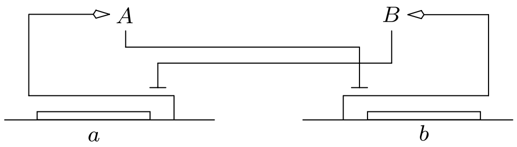

Fig. 1. A bistable switch. Expression of gene a (b) leads to synthesis of protein A (B) (arrow), which inhibits the expression of gene b (a) (dash).

The computational complexity of the identification of the whole system depends on the dimension n and on the number of approximation points. In our current implementation, for every fixed index i, the $\bar{x}^{(h)}$ and the $\delta^{(h)}$ are chosen so that sets $\mathscr{X}_i$ form a uniform partition of the (finite) domain of $x_{\ell(i)}$. However, we make this partitioning implicit and keep track only of those sets $\mathscr{X}_i$ that are "visited" by the data. The resulting procedure has complexity $O(n \times \max\{l(i)\} \times N_y)$, where N_y denotes the total number of data points. As a consequence, the worst-case complexity for the identification of the whole system is polynomial, namely $O(n^2 \times N_y)$. This remarkable feature was achieved by exploiting the quasi-diagonal structure of the model. As a matter of comparison, methods that do not exploit this structure — for instance, the clustering method presented in Drulhe et al. (2006) — typically have exponential complexity. As a final remark, note that the application of the algorithm to the i-th equation does not require the simultaneous observation of the whole state vector; it is only required that y_i and $y_{\ell(i)}$ be simultaneously available.

5. EXAMPLES

We will now demonstrate our identification procedure on two benchmark examples of genetic regulatory networks.

5.1 *Bistable switch*

This network is often found as a subsystem of actual regulatory networks (Farcot and Gouze (2006)), though here we take this as a standalone example without referring to any real system. The network is composed of two genes, say a and b, that both inhibit the expression of the other via the synthesis of the corresponding proteins A and B. A schematic view of the system is given in Figure 1. On a qualitative basis, depending on the initial concentrations and on perturbations (inputs), this system has two possible stable equilibria: high A and low B concentration, or low A and high B concentration. Let x_1 and x_2 denote the concentration of proteins A and B. The equations of this system with simplified notation are:

$$\begin{aligned} x_1(t+1) =& a_1 x_1(t) + b_1 u_1(t) + \bar{b}_1, \\ x_2(t+1) =& a_2 x_2(t) + b_2 u_2(t) + \bar{b}_2, \end{aligned}$$

where $p_1(x_2) = s^-(x_2; \eta_1, d_1)$ and $p_2(x_1) = s^-(x_1; \eta_2, d_2)$. Figure 2 shows the evolution of the state starting from different initial conditions. The trajectories were generated at random according to the model with the parameter values reported in Table 1. These values are somehow arbitrary and will be regarded as the true values of the system.

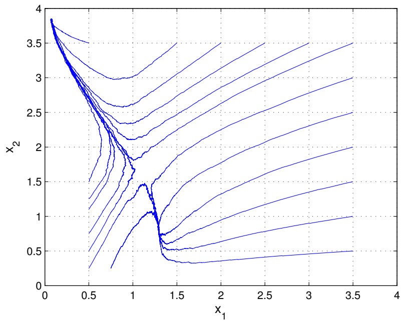

Fig. 2. Random evolution of the system from different initial conditions.

a_1	b_1	$\bar{b}_1$	η_1	d_1
0.998	0.0025	0	2	5

a_2	b_2	$\bar{b}_2$	η_2	d_2
0.999	0.0037	0	1	5

Table 1. Parameter values for the bistable switch model.

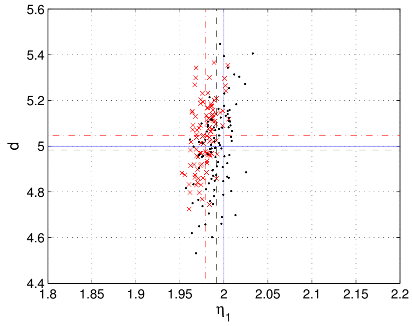

Fig. 3. Scatter plots of 100 parameter estimates, for $N = 1$ (black dots) and $N = 10$ (red crosses). Solid lines: true parameter values. Black dashed ($N = 1$) and red dash-dotted ($N = 10$) lines: mean estimate values.

We shall now consider identification of parameters η_1 and d_1 pertaining to the dynamics of x_1. Given the symmetry of the system, the procedure for the identification of parameters η_2 and d_2 is identical. We consider the observations of 60 simulated state trajectories starting from the point $(x_1, x_2) = (0.1, 0.2)$ (where both genes are on with high probability). The standard deviation of the Gaussian measurement noise was set to $\sigma_1 = \sigma_2 = 0.04$. We considered equally spaced observations for the two different observation rates $N = 1$ and $N = 10$, for a total number of observations per trajectory equal to 3000 and 300, respectively. The approximation domains $\mathscr{X}_1$ were chosen so as to partition the state space into stripes of width $\delta_2^{(h)} = 0.125$. In order to improve the robustness of the estimation, we discarded those partitions which were explored by less than 15 trajectories.

Scatter plots of the results from 100 repetitions of the estimation procedure are reported in Figure 3. Mean and variance of the estimates are reported in Table 2. The agreement of the parameter estimates with the true parameters is fairly good, however a small bias is present and becomes larger as the observations become sparser.

	$\hat{\eta}_1(2)$	$\hat{d}(5)$	$\hat{\eta}_1(2)$	$\hat{d}(5)$
mean	1.9913	4.9829	1.9788	5.0477
var	0.0002	0.0341	0.0001	0.0209

Table 2. Mean and variance of the parameter estimates. Left: $N = 1$; Right: $N = 10$. True parameter values are reported in brackets.

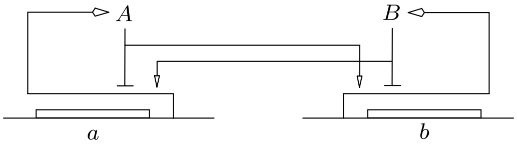

Fig. 4. Bistable system with autoregulation. Expression of gene a (b) leads to synthesis of protein A (B) (horizontal arrow), which promotes the expression of gene b (a) (vertical arrow) and inhibits the expression of gene a (b) (dash).

a_1	b_1	$\bar{b}_1$	$\eta_{1,1}$	$d_{1,1}$	$\eta_{1,2}$	$d_{1,2}$
0.998	0.02	0	8	4	4	4

a_2	b_2	$\bar{b}_2$	$\eta_{2,1}$	$d_{2,1}$	$\eta_{2,2}$	$d_{2,2}$
0.998	0.02	0	8	4	4	4

Table 3. Parameter values for the bistable system with autoregulation.

This is possibly due to the effects of noise, which may alter the assignment of data points to the correct partitions of the state space. Bias could also be due to a systematically unbalanced spreading of the observations within the partitions, leading to empirical estimates $\hat{v}_i(\bar{x})$ that are effectively computed about a point $x \in \mathscr{X}_i(\bar{x}, \delta)$ away from $\bar{x}$. This raises a tradeoff in the choice of the tolerance vector δ; smaller tolerance guarantees better localization but reduces the amount of observations in the approximation region.

5.2 A bistable two-state system with autoregulation

This system consists of two genes, each of which upregulates the expression of the other and downregulates its own expression. Since both thresholds related to self-repression are higher than the thresholds related to cross-activation, the system comprises a positive feedback circuit with two possible steady states: both genes on, and both genes off (Thomas and Kaufman (2001)). The structure of the network is reported in Figure 4. Let x_1 and x_2 denote the concentration of proteins A and B. The equations of the system are:

$$\begin{aligned} x_1(t+1) =& a_1 x_1(t) + b_1 u_{1,1}(t) u_{1,2}(t) + \bar{b}_1, \\ x_2(t+1) =& a_2 x_2(t) + b_2 u_{2,1}(t) u_{2,2}(t) + \bar{b}_2, \end{aligned}$$

with transition probabilities

$$\begin{aligned} p_{1,1}(x_1) =& s^-(x_1; \eta_{1,1}, d_{1,1}), \quad p_{1,2}(x_2) = s^+(x_2; \eta_{1,2}, d_{1,2}), \\ p_{2,1}(x_1) =& s^+(x_1; \eta_{2,1}, d_{2,1}), \quad p_{2,2}(x_2) = s^-(x_2; \eta_{2,2}, d_{2,2}). \end{aligned}$$

Figure 5 shows the evolution of the state starting from different initial conditions. The trajectories were generated at random according to the model and the parameter values reported in Table 3. These values will be regarded as the true values of the system.

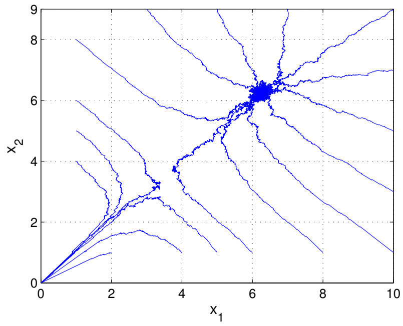

Fig. 5. Random evolution of the system from different initial conditions.

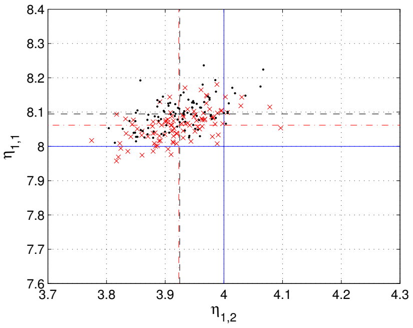

Fig. 6. Scatter plots of 100 parameter estimates, for $N = 1$ (black dots) and $N = 10$ (red crosses). Solid lines: true parameter values. Black dashed ($N = 1$) and red dash-dotted ($N = 10$) lines: mean estimate values.

	$\hat{\eta}_{1,2}(4)$	$\hat{\eta}_{1,1}(8)$	$\hat{d}(4)$	$\hat{\eta}_{1,2}(4)$	$\hat{\eta}_{1,1}(8)$	$\hat{d}(4)$
mean	3.924	8.094	3.765	3.923	8.061	3.875
var	0.002	0.002	0.003	0.003	0.002	0.004

Table 4. Mean and variance of the parameter estimates. Left: $N = 1$; Right: $N = 10$. True parameter values are reported in brackets.

As in the previous example, the system is symmetric, therefore we will limit ourselves to the identification of the parameters $\eta_{1,1}$, $d_{1,1}$, $\eta_{1,2}$ and $d_{1,2}$ relevant to the first state equation. In this example, we assume that $d_{1,1} = d_{1,2} = d$ and consider estimation of d, $\eta_{1,1}$ and $\eta_{1,2}$ given the observations of 60 simulated state trajectories starting from $(x_1, x_2) = (12, 10)$ and of additional 60 starting from $(x_1, x_2) = (9, 2)$. The standard deviation of the Gaussian measurement noise was set to $\sigma_1 = \sigma_2 = 0.02$. We considered equally spaced observations for the two different observation rates $N = 1$ and $N = 10$, for a total number of observations per trajectory equal to 5000 and 500, respectively. The approximation domains $\mathscr{X}_1$ were chosen so as to partition the state space into squares with edge size 0.8. To improve robustness of estimation, we discarded those partitions which were explored by less than 30 trajectories.

Scatter plots of the results from 100 repetitions of the estimation procedure are reported in Figure 6 (for better visualization, estimates of d were not included in the plot). Mean and variance of the estimates are reported in Table 4. The parameter estimates are more biased than in the previous example, but there is no significant difference between the results for $N = 1$ and $N = 10$. Therefore we argue that, in this case, the limiting factors are the observation noise and the distribution of the data in the state space. In particular, the greater complexity of this system (because of the autoregulation mechanisms both equations depend on both states) makes the estimation procedure more sensitive to the initial values of the state of the observed trajectories. Intuitively speaking, depending on the initial state, the observed trajectories most likely cross two, one, or none of the thresholds that characterize the stochastic laws of the product $u_{1,1}u_{1,2}$, with clear impact on the richness of the data set. Although we did not investigate optimal experimental design, we discovered that the use of two appropriately placed initial conditions significantly improves the estimation quality (note that, in our method, the initial conditions for the various experiments need not be the same). As a general guideline, we suggest that diversifying the data (e.g. by starting biological experiments from several different initial conditions), is an effective way to improve the estimation accuracy.

6. CONCLUDING REMARKS

We presented a stochastic hybrid framework for the description of genetic regulatory networks that is a natural extension of the well-known sigmoidal and piecewise affine models. Based on this framework, we have devised a procedure for the identification of some key parameters of the model that can be generalized to the identification of all parameters. The procedure exploits the structure of the model by means of local approximations of the stochastic dynamics. It does not demand that all observed trajectories start from the same initial conditions. In addition, not all components of the state need to be observed simultaneously. The estimation performance was demonstrated on simple benchmark systems. Several refinements of the identification procedure are already envisioned: 1) the adaptive choice of $\bar{x}^{(h)}$ and δ^h on the basis of the distribution of the observations over the state space; 2) the use, for every given point $\bar{x}^{(h)}$ and index i, of several non-overlapping data portions from the same observed trajectory; 3) the development of ad-hoc methods for the solution of the optimization problem (13); 4) the extension of the estimation procedure to all model parameters. Finally, application of our identification procedure to a stochastic hybrid model of *Escherichia Coli* carbon starvation response is being considered.

REFERENCES

R. Alur, C. Belta, F. Ivancic, V. Kumar, M. Mintz, G. Pappas, H. Rubin, and J. Schug. Hybrid modeling and simulation of biological systems. In M. Di Benedetto and A. Sangiovanni-Vincentelli, editors, *Hybrid Systems: Computation and Control*, number 2034 in LNCS, pages 19–32, Berlin, 2001. Springer–Verlag.

G. Batt, D. Ropers, H. de Jong, J. Geiselmann, R. Mateescu, M. Page, and D. Schneider. Validation of qualitative models of genetic regulatory networks by model checking: Analysis of the nutritional stress response in escherichia coli. *Bioinformatics*, 21(1):i19–i28, 2005.

E. Cinquemani, R. Porreca, G. Ferrari-Trecate, and J. Lygeros. Parameter identification for stochastic hybrid models of biological interaction networks. In *Proceedings of the 46th IEEE Conference on Decision and Control*, 2007.

H. de Jong. Modeling and simulation of genetic regulatory systems: A literature review. *Journal of Computational Biology*, 9(1):69–105, 2002.

H. de Jong, J.-L. Gouze, C. Hernandez, M. Page, T. Sari, and J. Geiselmann. Hybrid modeling and simulation of genetic regulatory networks: A qualitative approach. In O. Maler and A. Pnueli, editors, *Hybrid Systems: Computation and Control*, number 2623 in LNCS, pages 267–282, Berlin, 2003. Springer–Verlag.

S. Drulhe, G. Ferrari-Trecate, H. de Jong, and A. Viari. Reconstruction of switching thresholds in piecewise-affne models of genetic regulatory networks. In J. Hespanha and A. Tiwari, editors, *Hybrid Systems: Computation and Control*, number 3927 in LNCS, pages 184–199, Berlin, 2006. Springer–Verlag.

M.J. Dunlop, E. Franco, and R. M. Murray. A multi-model approach to identification of biosynthetic pathways. In *Proceedings of the 26th American Control Conference*, 2007.

E. Farcot and J.L. Gouze. How to control a biological switch: a mathematical framework for the control of piecewise affine models of gene networks. Technical Report 5979, INRIA, Sophia Antipolis Cedex (France), September 2006.

K. Fujarewicz, M. Kimmel, and A. Swierniak. On fitting of mathematical models of cell signaling pathways using adjoint systems. *Mathematical Biosciences and Engineering*, 2(3):527–534, 2005.

R. Ghosh and C.J. Tomlin. Symbolic reachable set computation of piecewise affine hybrid automata and its application to biological modeling: Delta-notch protein signaling. *IET Systems Biology*, 1(1):170–183, 2004.

J. Hu, W.C. Wu, and S.S. Sastry. Modeling subtilin production in *bacillus subtilis* using stochastic hybrid systems. In R. Alur and G.J. Pappas, editors, *Hybrid Systems: Computation and Control*, number 2993 in LNCS, pages 417–431, Berlin, 2004. Springer–Verlag.

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

P. Kouretas, K. Koutroumpas, J. Lygeros, and Z. Lygerou. Stochastic hybrid modeling of biochemical processes. In C. G. Cassandras and J. Lygeros, editors, *Stochastic Hybrid Systems*, volume 24 of *Automation and Control Engineering Series*. CRC press, 2006.

H.H. McAdams and A. Arkin. Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences, USA*, 94:814–819, 1997.

H.H. McAdams and A. Arkin. It's a noisy business! genetic regulation at the nanomolar scale. *Trends in genetics*, 15(2):65–69, 1999.

T.J. Perkins, M. Hallett, and L. Glass. Inferring models of gene expression dynamics. *Journal of Theoretical Biology*, 230(3):289–299, 2004.

R. Thomas and M. Kaufman. Multistationarity, the basis of cell differentiation and memory. I. structural conditions of multistationarity and other nontrivial behavior. *Chaos*, 11(1):170–179, 2001.

J.M. Vilar, H.Y. Kueh, N. Barkai, and S. Leibler. Mechanisms of noise-resistance in genetic oscillators. *PNAS*, 99:5988–5992, 2002.