Parameter identification for stochastic hybrid models of biological interaction networks

Eugenio Cinquemani, Riccardo Porreca, Giancarlo Ferrari-Trecate, and John Lygeros

Abstract—Based on a model of subtilin production by Bacillus subtilis, in this paper we discuss the parameter identification of stochastic hybrid dynamics that are typically found in biological regulatory networks. In accordance with the structure of the model, identification is split in two subproblems: estimation of the genetic network regulating subtilin production from gene expression data, and estimation of population dynamics based on nutrient and population profiles. Techniques for parameter estimation from sparse and irregularly sampled observations are developed and applied to simulated data. Numerical results are provided to show the effectiveness of our methods.

I. INTRODUCTION

During the last decades, different approaches to modeling of biochemical networks have been suggested in the literature [8]. Recently there has been a growing interest in the application of hybrid systems techniques to biological modeling and analysis [1], [9], [2], [10], [4], [13], [5]. It is also being recognized that many biological processes are intrinsically uncertain [25], [18], and stochastic phenomena appear to be instrumental for certain biochemical processes to improve robustness [27] or induce variability [28], and to play a key role in fundamental processes such as DNA replication [22]. Including stochasticity in the modeling framework leads to the development of stochastic hybrid models of diverse biological processes [19], [14].

Literature on hybrid biological model identification is only now beginning to appear [11], [23]. In [10], [24], efforts were devoted to specialize general hybrid systems identification techniques to the typical structure of genetic regulatory mechanisms, with clear advantages in terms of the quality of the reconstructed models.

In this paper we concentrate on the identification of the system that regulates the synthesis of the antibiotic subtilin by Bacillus subtilis. A stochastic hybrid model for this system was proposed in [16]. In this model, the continuous dynamics of the system depend on the discrete state of a genetic regulatory network. Switches among discrete state values are described in terms of both deterministic and stochastic laws. In [19], the model was further elaborated and cast into the framework of piecewise deterministic Markov processes [7]. The approach to identification that we follow is driven by the specific application and gives rise to a number of estimation procedures that may at first appear restricted to the particular system at hand. However, the different challenges posed by the identification of the subtilin production model are common to many biological identification problems:

- **Measurements at different levels of the biological scale.** Biological systems are studied at different levels of abstraction, varying from molecules to species. These levels are interacting and models often span more than one level. As an example, in the biological system considered, equations describing the dynamics of a population of cells and the level of global nutrient are coupled with the dynamics of gene expression that takes place in single cells. Data pertaining to different levels of abstraction are generally measured by means of different experiments and have different features. In our work the issue of correlating concentration and population data is explicitly taken into account.
- **Irregular sampling.** In biological experiments, it is often difficult to obtain regularly sampled data. This calls for identification techniques capable to cope with arbitrary sampling.
- **Stochastic variability.** Uncertainty is often seen as an additional difficulty of biological identification. A classical approach is to eliminate data uncertainty by averaging. This can be a wasteful operation in that it reduces the amount of information available. To make better use of probabilistic data, stochastic modeling is essential.

We will take the work in [16] as a starting point and focus on estimation of the model parameters. A preliminary study of identification for *B. subtilis* has been carried out in [20] by randomized methods, in a case where a subset of the continuous state variables of the system is observed. In analogy with [10], we specialize identification methods for piecewise affine systems (see e.g. [12], and references therein) to the identification of deterministically switching gene dynamics. Moreover, new techniques for the estimation of the dynamics and of the switching probabilities of Markovian gene expression are introduced. Finally, a modification of the prediction error identification method [21], [26] is applied to the parametric identification of nonlinear switching dynamics driven by stochastic inputs.

The paper is organized as follows. In Section II, the *B. subtilis* subtilin production model is reviewed and the identification problem is stated. Section III describes tech-
niques for estimating cell-level parameters of the model from observations of the gene expression profiles. Section IV deals with estimation of population-level parameters based on macroscopic observations such as population size and nutrient level. A short section (Section V) is dedicated to the estimation of the threshold on the nutrient level on which the activation of subtilin production depends. Simulation results provided in Section VI show the effectiveness of our identification techniques. Conclusions and perspectives of our work are given in Section VII.

II. PRODUCTION OF SUBTILIN

Subtilin is an antibiotic synthesized by B. subtilis as an adaptive response to changes in the environment, allowing the cell to benefit optimally from the available resources. When the amount of nutrients is sufficient, B. subtilis population increases without a remarkable change in subtilin concentration. Subtilin production starts when the amount of nutrient falls under a threshold because of excessive population growth. The role of subtilin is to increase food supply by eliminating competing species and/or other B. subtilis cells. In addition to reducing the demand for nutrients, the decomposition of the cells killed by subtilin releases additional nutrients in the environment. The biosynthesis of subtilin is regulated by a positive feedback mechanism in which extracellular subtilin activates the two components regulatory system SpaK and SpaR that binds to a DNA motif promoting the expression of genes for subtilin biosynthesis (spaS and spaBTC) and immunity (spaIFEG). SpaK and SpaR react to form the complex SpaRK that will be used in our model. SpaRK expression is controlled by the sporulation transcription factor SigH. Finally, the composition of SigH is turned on whenever the nutrient concentration falls below a certain threshold. In this paper a simplified model is examined, in which spaBTC and spaIFEG are not taken into consideration.

A. Model

We refer to the dynamic model of subtilin production that was originally proposed in [16]. In this model, normalized population level $x_1$, nutrient level $x_2$ and the three concentrations $x_3 = \text{[SigH]}, x_4 = \text{[SpaRK]}$ and $x_5 = \text{[SpaS]}$ constitute the (nonnegative) continuous part $x = [x_1\ x_2\ x_3\ x_4\ x_5]^T$ of the system state. The discrete part of the state, $S = \{S_3, S_4, S_5\}$, is composed of three binary switches that account for whether the expression of SigH, spaRK and spaS genes (in this order) is activated ($S_i = 1$) or inhibited ($S_i = 0$).

The growth of B. subtilis population is governed by a logistic-type equation:

$$\dot{x}_1 = rx_1 (1 - x_1/D_\infty(x_2)).$$

(1)

For fixed $D_\infty$, the equilibrium point $x_1 = D_\infty$ is also the asymptotic value of $x_1$, provided $x_1(0) > 0$. The limiting population is assumed to depend on the nutrient level according to

$$D_\infty(x_2) = \min \{x_2/X_0, D_{\max}\},$$

(2)

where constants $X_0 > 0$ and $D_{\max} > 0$ reflect certain properties of the experimental environment. Nutrient consumption and production are governed by the equation

$$\dot{x}_2 = -k_1x_1 + k_2x_5,$$

(3)

where $k_1$ and $k_2$ are the rate of nutrient consumption per unit of population and the rate of increase in nutrient availability due to the action of subtilin, respectively. The remaining continuous states follow first-order linear dynamics depending on the current value of the corresponding discrete state $S_i$:

$$\dot{x}_i(t) = \begin{cases} -l_ix_i(t), & S_i(t) = 0, \\ -l_ix_i(t) + k_i, & S_i(t) = 1, \end{cases} \quad i = 3, 4, 5$$

(4)

where $l_i$ and $k_i$ are the natural degradation and synthesis rates of the corresponding protein complexes. The rule that describes the status of SigH production is deterministic and given by

$$S_3(t) = \begin{cases} 1, & x_2(t) < \eta; \\ 0, & x_2(t) \geq \eta, \end{cases}$$

(5)

where threshold $\eta$ indicates the nutrient level below which subtilin production mechanism is triggered. Finally, switches $S_4$ and $S_5$ are modeled as binary random processes, which makes the system a stochastic hybrid one. In [16], a discrete-time model of $S_4$ and $S_5$ is expressed in terms of the following switching probabilities: for $i = 4$ and $i = 5$,

$$\mathbb{P}[S_i(kT + T) = 1|S_i(kT) = 0] = a_0(x_{i-1}(kT)) = \frac{c_ix_{i-1}(kT)}{1 + c_ix_{i-1}(kT)};$$

$$\mathbb{P}[S_i(kT + T) = 0|S_i(kT) = 1] = a_1(x_{i-1}(kT)) = \frac{1}{1 + c_ix_{i-1}(kT)}.$$

(6)

(7)

The implicit assumption is that $S_4$ and $S_5$ can be considered constant between samples:

$$S_i(t) = S_i(kT), \quad \forall t \in [kT, kT + T).$$

(8)

The fact that $a_0 = 1 - a_1$ has the following important consequence.

**Proposition I:** If $a_0 = 1 - a_1$ then $\{S_i(kT) : k = 0, 1, \ldots\}$ is a sequence of mutually independent random variables.

From now on, for simplicity, we shall denote $S_i(kT)$ and $x_i(kT)$ by $S_i(k)$ and $x_i(k)$, respectively. For given values of $x_3(k)$ and $x_4(k)$, $S_4$ and $S_5$ are assumed to be conditionally independent and Markovian in the following sense:

$$p(S_4(k+1), S_5(k+1)|S_4^-(k), S_5^-(k), x_3(k), x_4(k)) = p(S_4(k+1)|S_4^-(k), x_3(k))p(S_5(k+1)|S_5^-(k), x_4(k))$$

(9)

where $S_i^-(k) = \{S_i(\ell) : \ell \leq k\}$ and $p(\cdot)$ denotes conditional probability distribution.

This model of subtilin production can be thought of as the interconnection of two deterministic dynamical subsystems with stochastic inputs, as depicted in Fig. 1. Subsystem $M$...
is composed of equations (1) and (3), whereas subsystem \( \mu \) implements the cascade of switching linear systems (4). Feedback interconnection is operated by two static subsystems: gain \( k_2 \) and threshold \( \eta \). Process \( v \) is a fictitious stochastic input that governs the stochastic switching of \( S_4 \) and \( S_5 \). From a biological point of view, this decomposition reflects two different levels in the scale of biological abstraction. Model \( \mathcal{M} \) describes the evolution of macroscopic variables such as nutrient and population level. Microscopic phenomena such as (average) gene expression are accounted for by model \( \mu \). The interface between microscopic and macroscopic levels is formalized by converting the amount of available nutrient \( (x_2) \) into a triggering signal for the genetic expression mechanism \((S_3)\) and, conversely, by relating the expression level of the SpaS gene to the production of new nutrient.

Fig. 2 shows the trajectories generated by a typical run of the model. Simulation is carried out with sampling time \( T = 0.72 \text{sec} \). The parameter values used in the experiment are shown in Table I. Hereafter they will be regarded as the “true” parameter values. One may observe that, starting from abundance of nutrients, the nonzero initial population grows freely for about 300 minutes, consuming part of the nutrients available. When the level of nutrients falls below threshold \( \eta = 4 \), the subtilin production mechanism is triggered. As more subtilin is released, new nutrients become available. Conversely, when threshold \( \eta \) is crossed from below, subtilin production is switched off, and the nutrient level starts to decay. This explains the sustained oscillations of the nutrient level visible for \( t \geq 320 \text{min} \). Fast random switches of the dynamics of [SpaRK] and [SpaS] may be appreciated e.g. in the time interval \([700, 900] \text{min}\).

### TABLE I

<table>
<thead>
<tr>
<th>( r )</th>
<th>( X_0 )</th>
<th>( D_{\text{max}} )</th>
<th>( k_1 )</th>
<th>( k_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.02</td>
<td>4</td>
<td>1</td>
<td>0.1</td>
<td>0.4</td>
</tr>
<tr>
<td>( \eta )</td>
<td>( k_3 )</td>
<td>( l_3 )</td>
<td>( c_4 )</td>
<td>( k_4 )</td>
</tr>
<tr>
<td>4</td>
<td>0.5</td>
<td>0.2</td>
<td>1</td>
<td>0.2</td>
</tr>
</tbody>
</table>

**B. Discrete-time model**

We will always use discrete-time versions of the models of \( \mu \) and of \( \mathcal{M} \) sampled at multiples of \( T \). By assumption (8), the following discrete-time state-space model \( \mu \) is found by exact integration of (4):

\[
x_i(k + 1) = \begin{cases} \tilde{l}_i x_i(k), & S_i(k) = 0, \\
\tilde{l}_i x_i(k) + \tilde{k}_i, & S_i(k) = 1, \end{cases} \quad i = 3, 4, 5
\]

where \( \tilde{l}_i \triangleq e^{-\frac{r}{T_i}} \) and \( \tilde{k}_i \triangleq \frac{k_i}{T_i} (1 - \tilde{l}_i) \). On the other hand, integrating (1) (resp. (3)) under the approximation that \( x_2 \) (resp. \( x_1 \) and \( x_5 \)) is constant between samples yields the following discrete-time state-space model for \( \mathcal{M} \):

\[
\begin{bmatrix} x_1(k + 1) \\ x_2(k + 1) \end{bmatrix} = \begin{bmatrix} f(x_1(k), x_2(k)) \\ x_2(k) - \tilde{k}_1 x_1(k) \end{bmatrix} + \begin{bmatrix} 0 \\ \tilde{l}_2 \end{bmatrix} x_5(k),
\]

with \( \tilde{r} \triangleq e^{-\frac{r}{T_i}} \), \( \tilde{k}_1 = T k_1 \), \( \tilde{k}_2 = T k_2 \), and

\[
f(x_1, x_2) \triangleq D_{\text{max}}(x_2)/\left[ 1 + (D_{\text{max}}(x_2)/x_1 - 1) \tilde{r} \right].
\]

Note that the sampling period \( T \) is not related to the time elapsed between consecutive measurements but is just a modeling constant that can be chosen sufficiently small in order to make piecewise constant approximation of the state variables acceptable.

### C. Measurement model

We consider measurements of the population-level variables \( x_1, x_2 \) and of the cell-level variables \( x_3, x_4, x_5 \) separately. In other words, we allow systems \( \mathcal{M} \) and \( \mu \) to be observed in two different experiments (observation of nutrient and population levels vs. observation of gene expression levels) starting from the same initial conditions but with possibly different observation times.

**Definition 1:** The sequence \( Z = \{ Z_0, Z_1, \ldots \} \subseteq \mathbb{N} \) is a time scale if \( Z_\ell < Z_{\ell+1}, \forall \ell \in \mathbb{N} \). For a function \( g : Z \rightarrow \mathbb{R} \) and numbers \( a \in \mathbb{N}, b \in \mathbb{N} \) such that \( a < b \), \( g|_{[a, b]} \) denotes the set \( \{ g(Z_a), g(Z_{a+1}), \ldots, g(Z_b) \} \). If \( Z' \subseteq Z \) is a time scale, \( g|_{Z'} \) is the restriction of \( g \) to \( Z' \). \( Z = \mathbb{N} \) will be called the natural time scale.

Let \( T \) and \( \tau \) be two finite time scales. Measurements are modeled as follows:

\[
y_i(T_\ell) = x_i(T_\ell) + e_i(T_\ell), \quad i = 1, 2,
\]

\[
y_i(\tau_i) = x_i(\tau_i) + e_i(\tau_i), \quad i = 3, 4, 5,
\]
where the time scales $T$ and $\tau$ denote the observation instants in the two experiments, and the $e_i$ are mutually independent zero-mean white Gaussian measurement noise with known variance $\sigma_i^2$. Lags between consecutive measurements will be denoted by $\Delta t = T_{i+1} - T_i$ and $\delta t = \tau_{i+1} - \tau_i$.

D. Identification problem

Based on measurements (12)–(13), we want to estimate the parameters of system $\mathcal{M}$, i.e. $\theta_{\mathcal{M}} = (r, D_{\text{max}}, X_0, k_1)$, those of system $\mu$, i.e. $\theta_{\mu} = (k_3, l_3, k_4, l_4, c_4, k_5, l_5, c_5)$, and the coupling parameters $\eta$ and $k_2$. The idea is to split the identification problem into simpler subproblems:

(I) estimation of $\theta_{\mu}$ given $\mathcal{Y}_\mu \equiv \{\tilde{y}_3^i, \tilde{y}_4^i, \tilde{y}_5^i\}$;

(II) estimation of $\theta_{\mathcal{M}}$ given $\mathcal{Y}_\mathcal{M} = \{\tilde{y}_4^i, \tilde{y}_5^i\}$;

(III) estimation of the coupling parameters $\eta$ and $k_2$.

In principle, all the above problems are mutually dependent. For instance, the evolution of $\mu$ depends on state $x_2$ through the threshold $\eta$, but $x_2$ is not observed in problem (I). However, the only effect of $x_2$ on $\mu$ is to trigger the switches of $S_3$. Therefore, if $S_3^3$ was known, one could solve problem (I) without using observations of $x_2$. In order to decouple problem (I) from (II) and (III) we will apply a segmentation procedure to isolate the switches in the time series $\tilde{y}_3^i$. Similarly, system $\mathcal{M}$ is driven through gain $k_3$ by state $x_5$, which is not observed when problem (II) is considered. To cope with this, we shall assume that certain statistics of process $x_5$ are known. In practice, they can be derived separately from multiple experiments on $\mu$. This turns problem (II) into a standard nonlinear stochastic identification problem where $\theta_{\mathcal{M}}$ and $k_2$ are the unknown parameters. Finally, multiple experiments on $\mathcal{M}$ and $\mu$ can be used for the estimation of $\eta$. To summarize, we will proceed as follows:

a) estimation of $\theta_{\mu}$ given $\mathcal{Y}_\mu$ (Section III);

b) estimation of $\theta_{\mathcal{M}}$ and $k_2$ given $\mathcal{Y}_\mathcal{M}$ (Section IV);

c) estimation of $\eta$ given $\mathcal{Y}_\mu$ and $\mathcal{Y}_\mathcal{M}$ (Section V).

Given that the discrete-time model parameters $\tilde{r}$, $\tilde{k}_3$, and $\tilde{l}_3$ are in one-to-one correspondence with $r$, $l_3$, and $k_3$, when convenient, we will discuss estimation in terms of $\tilde{r}$, $\tilde{k}_3$, and $\tilde{l}_3$.

III. IDENTIFICATION OF $\mu$

The dynamics (10) has a cascade structure in the sense that, for $i = 3, 4, 5$, $S_i$ drives the dynamics of $x_i$ that in turns affects the discrete state $S_{i+1}$. We exploit this structure for splitting problem (a) into the separate identification of three switching linear systems of type (10) from the corresponding output measurements $y_i$. For $i = 3$, switches occur at a low rate (see Figure 2) and we expect several measurements to be available between them. On the other hand, for $i = 4$ and $i = 5$, the switching rate is typically much larger than the measurement rate. Identification of $\mu$ is therefore arranged in three subsequent steps, in which the identification procedure depends on the nature of the switching mechanism:

• $i = 3$: Based on measurements $y_3^i$, isolate portions of $S_3^3$ without switches. Within each portion, reconstruct the dynamics of $x_3^i$ using $\tilde{y}_3^i$. This yields estimates of $l_3$, $k_3$, and of $x_3^N$ (Section III-A);

• $i = 4, 5$: Based on the estimates of $x_{i-1}^N$, compute the probabilities of switching of $S_i^N$ as a function of $c_i$, and the expected values $\mathbb{E}[x_i^\ell]$ as a function of $c_i$, $l_i$, $k_i$. Next, identify the dynamics of $\mathbb{E}[x_i^\ell]$ using measurements $\tilde{y}_i^\ell$. This yields estimates of $c_i$, $l_i$, $k_i$, and $x_i^N$ (Section III-B).

A. Case $i = 3$

Throughout this section we shall rely on the following simplifying assumption.

Assumption 1: At most one switch of $S_3^3$ may occur between two consecutive samples of $y_3^i$.

This approximation is supported by the fact that multiple switches between consecutive observations are very seldom observed in the experimental settings of our concern.

For a fixed index $\ell$, consider the case where $S_3(\tau_\ell) = S_3(\tau_{\ell+1})$, i.e. $S_3^3|_{[\tau_\ell, \tau_{\ell+1}]}$ is constant. Then, one has:

$$x_3(\tau_{\ell+1}) = \begin{cases} L_3 x_3(\tau_\ell) + K_3, & \text{if } S_3(\tau_\ell) = 1 \\ L_3 x_3(\tau_\ell), & \text{if } S_3(\tau_\ell) = 0 \end{cases},$$

where $L_3 = \tilde{I}_3^N$ and $K_3 = \tilde{k}_3(1 - \tilde{I}_3^N)/(1 - \tilde{l}_3)$. Note that, for each value of $\ell$, constants $L_3$ and $K_3$ are in one-to-one correspondence with $l_3$ and $k_3$, and hence with $l_3$ and $k_3$. Based on model (14), we propose methods for:

• estimating $\tilde{S}_3^3$ given $l_3$ and $k_3$ (Section III-A.1);

• estimating $l_3$, $k_3$, and $x_3^N|_{[\tau_\ell, \tau_{\ell+1}]}$, given $S_3^3$ (Section III-A.2);

and combine them to derive an iterative algorithm for the joint estimation of $l_3$, $k_3$, and $S_3^3$ (Section III-A.3). Finally, a method for computing estimates of the whole $x_3^N$ trajectory, which will be used in the identification problems $i = 4, 5$, is described in Section III-A.4.

1) Estimation of $S_3(\tau_\ell)$ given $l_3$ and $k_3$: The method is based on a statistical test between hypotheses $H_0 : S_3(\tau_\ell) = 0$ and $H_1 : S_3(\tau_\ell) = 1$. In particular, as shown in [6], this test reduces to the problem of discriminating the mean of a Gaussian distribution. According to well-known results for this problem (see e.g. [15], [3]), an optimal and symmetric test between $H_0$ and $H_1$ leads to the following estimator of $S_3(\tau_\ell)$:

$$\hat{S}_3(\tau_\ell) = \begin{cases} 1, & \text{if } y_3(\tau_{\ell+1}) - L_3 y_3(\tau_\ell) \geq K_3/2; \\ 0, & \text{if } y_3(\tau_{\ell+1}) - L_3 y_3(\tau_\ell) < K_3/2. \end{cases}$$

2) Estimation of $l_3$ and $k_3$ given the $S_3$ sequence: Assume that $S_3^N|_{[m,n]}$ is a known and constant sequence of discrete states. The parameters $l_3$ and $k_3$ can be estimated over this window solving the nonlinear least squares problem:

$$\min_{k_3,l_3,x_3(\tau_m)} J_3, \quad J_3 = \sum_{\ell=m}^n (y_3(\tau_\ell) - x_3(\tau_\ell))^2,$$

subject to (14).

From the solution $\hat{k}_3$, $\hat{l}_3$, $\hat{x}_3(\tau_m)$ to (16) and from (14) one can also compute the estimates of $x_3^N|_{[m,n]}$ and approximations of the variances $\sigma_i^2 \triangleq \text{var}(l_3)$, $\sigma_k^2 \triangleq \text{var}(k_3)$ (see [21]).
Consider time-scales $\tau(1) \subset \tau(2) \subset \tau(3) \subset \tau$ that are consecutive (i.e. if $i < j$ then $z < z'$, $\forall z \in \tau(i), z' \in \tau(j)$) and such that $S_3^{\tau(i)}$ is constant, $\forall i \in \{1, \ldots, s\}$. By applying the previous procedure one can associate to each time-scale $\tau(i)$ the local estimates $\hat{l}(i)$, $k(i)$ and their variances $\sigma^2_1(i)$, $\sigma^2_2(i)$. Local estimates can be combined into global estimates $\hat{l}_3[j]$, having variance $\sigma^2_3[j]$, using an iterative procedure based on the initialization $\hat{l}_3[1] = \hat{l}_3(1)$, $\sigma^2_3[1] = \sigma^2_3(1)$ and the update formulae

$$\begin{align*}
\hat{l}_3[j + 1] &= \frac{\sigma^{-2}_3(j + 1)\hat{l}_3(j) + \sigma^{-2}_3[j]\hat{l}_3[j]}{\sigma^{-2}_3(j + 1) + \sigma^{-2}_3[j]} (17) \\
\sigma^2_3[j + 1] &= \frac{1}{\sigma^{-2}_3(j + 1) + \sigma^{-2}_3[j]} (18)
\end{align*}$$

The procedure is identical for $k_3$, except for the case $S_3^{\tau(i)} = 0$, when we set $\sigma^2_3(j) = +\infty$ for avoiding to update global estimate since $k_3$ is irrelevant.

3) Joint estimation: Let $\hat{l}_3[0]$ and $k_3[0]$ be initial guesses of the values of $l_3$ and $k_3$. The iterative estimation of $l_3$, $k_3$ and $S_3$ reads as follows.

- **Initialization:** Set $\ell = 0$ and $j = 0$. Recalling (15), $\hat{S}_3(\tau_0)$ can be estimated as

$$\hat{S}_3(\tau_0) = \begin{cases} 1, & \text{if } y_3(\tau_{\ell+1}) - \hat{L}_3[j]y_3(\tau_0) \geq \hat{K}_3[j]/2; \\ 0, & \text{if } y_3(\tau_{\ell+1}) - \hat{L}_3[j]y_3(\tau_0) < \hat{K}_3[j]/2,
\end{cases}$$

where $\hat{L}_3[j] = e^{-\delta t_l}[j]$, $\hat{K}_3[j] = \frac{\hat{l}_3[j]}{l_3[0]} (1 - \hat{L}_3[j])$.

- **First step:** For $\ell \geq 0$ estimate $\hat{S}_3(\tau_{\ell+1})$ according to (20) until a time $\tau_{\ell_0}$ where $\hat{S}_3(\tau_{\ell_0})$ differs from $S_3^\tau$.

At this point define $\tau(1) = \{\tau_0, \ldots, \tau_{\ell_0-1}\}$, compute $\hat{x}_3(\tau_0)$, $\hat{l}_3(1)$, $k_3(1)$ along with their variances as described above and set $\hat{l}_3[1] = \hat{l}_3(1)$, $k_3[1] = k_3(1)$.

- **Main iteration:** Increment counter $j$. For $\ell \geq \ell_1$ estimate $\hat{S}_3(\tau_{\ell+1})$ from (19) until a time $\tau_{\ell+1}$ where $\hat{S}_3(\tau_{\ell+1})$ differs from $S_3^\tau$ or until $\tau_{\ell+1} = \tau_{\ell_{\max}}$ ($\tau_{\ell_{\max}}$ denotes the last element of $\tau$). Define $\tau(j + 1) = \{\tau_{\ell_1}, \ldots, \tau_{\ell_{\ell+1}}\}$, compute $\hat{x}_3(\tau_{j+1})$, $\hat{l}_3(j + 1)$, $k_3(j + 1)$ along with their variances and compute $\hat{l}_3[j+1]$, $k_3[j+1]$, $\sigma^2_3[j+1]$, $\sigma^2_3[j+1]$ according to (17)–(18). Repeat this step until all data have been processed and let $j_{\max} = \ell + 1$.

- **State estimation:** In view of (14), on each time-scale $\tau(j)$ the states $\hat{x}^{\tau(j)}$ can be estimated by running the model

$$\hat{x}_3(\tau_{\ell+1}) = \begin{cases} \hat{L}_3[j_{\max}]\hat{x}_3(\tau_{\ell+1}) + \hat{K}_3[j_{\max}], & \text{if } \hat{S}_3(\tau_{\ell+1}) = 1 \\ \hat{L}_3[j_{\max}]\hat{x}_3(\tau_{\ell+1}), & \text{if } \hat{S}_3(\tau_{\ell+1}) = 0
\end{cases}$$

where the states at the beginning of each time-scale are those computed in the first step and in the main iteration. Further details about the choice of initial parameter guesses $\hat{l}_3[0]$, $k_3[0]$ can be found in [6].

4) Refinement of the $x_3$ estimates: Using the algorithm described in Section III-A.3 one obtains the estimated states $\hat{x}_3^\tau$. However, for the identification of the $x_4$ and $x_3$ models, estimates $\hat{x}_3^{\tau_{\max}}$ are needed. In view of Assumption 1, if $\hat{S}_3(\tau_{\ell+1}) = \hat{S}_3(\tau_{\ell+1})$, then $\hat{x}_3^{\tau_{\ell+1}}$ can be obtained by simulating model (10) with initial state $\hat{x}_3(\tau_{\ell+1})$ and $\hat{l}_3 = \hat{l}_3(\tau_{\ell+1})$.

B. Case $i = 4, 5$

Consider the equation:

$$x(k + 1) = \begin{cases} \hat{L}_x(k), & S(k) = 0; \\ \hat{L}_x(k) + \hat{k}_i, & S(k) = 1, \end{cases}$$

and define $p_i(k) = \mathbb{P}[S(k) = i], i = 1, 2$. According to (6)–(7) and Proposition 1, $p_i(k)$ are given by

$$p_0(k) = a_1(z(k - 1)), \quad p_1(k) = a_0(z(k - 1)).$$

System (20) models the equations of $x_4$ (resp. $x_5$), with $z = x_3$ (resp. $z = x_4$). Assume that $z^N$ is known, so that $p_0$ and $p_1$ are known functions of $c$. We are interested in estimating parameters $l, k$ and $c$ based on measurements $\hat{y}_{m,n}$ of $x_{m,n}$, where $m$ and $n$ verify $n > m$ and define the observation window over which estimation is performed.

**Proposition 2:** For any $\ell = m, \ldots, n - 1$ and $k = 0, 1, \ldots, \delta$, define

$$u(\tau, k) = \sum_{j=0}^{k-1} j^{\ell-1-j}p_1(\tau + j).$$

If the sequence $u(\tau, k)$ is computed as

$$u(\tau, k) = p_1(\tau + k - 1) + \hat{I}_u(\tau, k - 1)$$

with $u(\tau, 0) = 0$, then it holds

$$\mathbb{E}[x(\tau_{\ell+1})] = \hat{I}_u \mathbb{E}[x(\tau_{\ell})] + \hat{I}_u(\tau, \delta),$$

where (22) is initialized with $\mathbb{E}[x(\tau_0)]$. The proof of Proposition 2 can be found in [6]. Notice that the value of $z(\tau_m - 1)$ is needed to compute $p_1(\tau_m)$, but $\tau_m - 1$ lies outside the observation window. A simple solution is to use the approximation $z(\tau_m - 1) = z(\tau_m)$. The idea is to use $\mathbb{E}[x(\tau_{\ell+1})]$ as an approximation of $x(\tau_{\ell+1})$ and to fit it to the data. For fixed values of $\ell$, $k$ and $c$ and initial condition $\mathbb{E}[x(\tau_m)]$, the averages $\mathbb{E}[x(\tau_{\ell+1})]$ are computed iteratively by the formulae in Proposition 2. Given measurements $\hat{y}_{m,n}$, estimates $\hat{l}, \hat{k}, \hat{c}$, and $\hat{E}[x(\tau_m)]$ are then computed as follows:

$$\min_{l, k, c, \hat{E}[x(\tau_m)]} J_{4,5, \ell}, \quad J_{4,5, \ell} = \frac{1}{2} \sum_{k=0}^{n} (y_\ell(k) - \mathbb{E}[x(\tau_{\ell+1})])^2.$$

The above method applies quite easily to the equation of $x_4$ by setting $\hat{x}_3^{\tau_{\max}} = \hat{x}_3^{\tau_{\max}}$, where $\hat{x}_3^{\tau_{\max}}$ can be
computed with the algorithms of Section III-A.4. Similarly, application of the method to the equation of $x_5$ requires the knowledge of the estimates $\hat{x}_5^{[\tau_m, \tau_n]}$. Based on the estimates $\hat{t}_4 = e^{-i\tau}T$, $\hat{t}_3 = \frac{k_1}{k_2}(1 - \hat{t}_4)$, $\hat{c}_4$ and $\mathbb{E}[x_4(\tau_m)]$ drawn from (23), estimates $\hat{x}_4^{[\tau_m, \tau_n]} = \mathbb{E}[x_4^{[\tau_m, \tau_n]}]$ are computed by running the equation

$$\hat{\mathbb{E}}[x_4(k + 1)] = \hat{t}_4 \hat{\mathbb{E}}[x_4(k)] + \hat{k}_4 \hat{p}_1(k)$$

for $k = \tau_m, \tau_m + 1, \ldots, \tau_n - 1$.

IV. IDENTIFICATION OF $\mathcal{M}$

We now turn to the identification of parameters $\theta_\mathcal{M}$ and $k_2$ from measurements $y_\mathcal{M}$. Recall that process $x_5$, which is the input of subsystem $\mathcal{M}$, cannot be measured at this stage. The definition of switching probabilities given in Section II makes $x_5(k)$ an uncorrelated process (see Proposition 1). Let us assume that its time-varying mean $\bar{x}_5(k)$ and variance $\text{var}(x_5(k))$ are known (in practice, they can be estimated from multiple experiments on $\mu$). We formulate identification of $\theta_\mathcal{M}$ and $k_2$ as the following optimization problem:

$$\min_{\theta_\mathcal{M}, k_2} J_{1,2}(\theta_\mathcal{M}, k_2), \quad J_{1,2} = \sum_{\ell=0}^{n} \left| y_{1,2}(T_\ell) - \hat{x}_{1,2}(T_\ell) \right|^2,$$

(24)

where $x_{1,2} = [x_1, x_2]^T$, $y_{1,2} = [y_1, y_2]^T$ and $\hat{x}_{1,2}(T_\ell)$ is an estimate of $x_{1,2}(T_\ell)$ depending on the values of $\theta_\mathcal{M}$ and $k_2$. In the context of linear stochastic systems, term $\hat{x}_{1,2}(T_\ell)$ is typically the Kalman predictor associated to the candidate parameter values. This is called Prediction Error Method (see e.g. [26]), which is proven to be asymptotically consistent. In light of this property, we wish to use a similar approach to the identification of $\theta_\mathcal{M}$ and $k_2$. Since $\mathcal{M}$ is nonlinear, we shall write $\hat{x}_{1,2}(T_\ell)$ as the extended Kalman predictor $\hat{x}_{1,2}(T_\ell|\ell - 1)$ associated to dynamics (11) and measurements (12). For any value $\bar{x}_1 = [\bar{x}_1, \bar{x}_2]^T$ of $x_{1,2}$, let

$$F(\bar{x}_{1,2}) = \frac{\partial}{\partial \bar{x}_{1,2}} \left[ f(x_{1,2}) \right]_{x_{1,2}} = \begin{bmatrix} \frac{f(x_{1,2})}{x_1} & x_0(1 - \tau)\frac{f(x_{1,2})}{x_2} \\ -k_1 & 1 \end{bmatrix}, \quad \frac{\bar{x}_2}{\bar{x}_1} \leq D_{\text{max}};$$

$$= \begin{bmatrix} \frac{f(x_{1,2})}{x_1} & 0 \\ -k_1 & 1 \end{bmatrix}, \quad \frac{\bar{x}_2}{\bar{x}_1} \geq D_{\text{max}}.$$

The extended Kalman estimator, $\hat{x}_{1,2}(k|\ell)$, of $x_{1,2}(k)$ given measurements $\{y_1^{[0, \ell]}, y_2^{[0, \ell]}\}$ and the (approximate) covariance matrix of the estimation error $x_{1,2}(k) - \hat{x}_{1,2}(k|\ell)$ obey the following recursion:

- Measurement update:
  $$\Sigma(T_\ell|\ell) = [I - K(\ell)]\Sigma(T_\ell|\ell - 1) [I - K(\ell)]^T$$
  $$+ K(\ell)\text{diag}(\sigma_1^2, \sigma_2^2) K(\ell)^T,$$
  $$\hat{x}_{1,2}(T_\ell|\ell) = \hat{x}_{1,2}(T_\ell|\ell - 1) + K(\ell)[y_{1,2}(T_\ell) - \hat{x}_{1,2}(T_\ell|\ell - 1)],$$

with $K(\ell) = \Sigma(T_\ell|\ell - 1)\Sigma(T_\ell|\ell - 1) + \text{diag}(\sigma_1^2, \sigma_2^2)^{-1}$;

- Time update: for $k = T_\ell, T_\ell + 1, \ldots, T_{\ell+1} - 1$,
  $$\Sigma(k + 1|\ell) = F(\hat{x}_{1,2}(k|\ell)) \Sigma(k|\ell) F(\hat{x}_{1,2}(k|\ell))^T$$
  $$+ \text{diag}(0, \text{var}(x_5(k)))$$
  $$\hat{x}_{1,2}(k + 1|\ell) = f(\hat{x}_{1,2}(k|\ell)) + [0, \hat{k}_2]^T \bar{x}_5(k).$$

The recursion is initialized by $\hat{x}_{1,2}(0|0) = \mathbb{E}[x_{1,2}]$ and $\Sigma(0|0) = \text{Var}(x_{1,2})$. The classical derivation of the extended Kalman filter can be found e.g. in [17]. Our version is a straightforward adaptation to the discrete-time case with $\Delta_t$-steps-ahead prediction. The expression of $F$ is discontinuous due to the presence of term $D_{\infty}(x_2) = \min\{x_2/X_0, D_{\text{max}}\}$ in $f$. The above algorithm allows one to evaluate the loss functional $J_{1,2}(\theta_\mathcal{M}, k_2)$ for arbitrary candidate parameter values. The optimization problem (24) can be solved by standard numerical methods.

V. IDENTIFICATION OF $\eta$

An estimate of $\eta$ may be drawn by the use of empirical statistics from multiple experiments on $\mu$. Our approach is based on the following observation: at any time $k$,

$$\mathbb{E}[S_3(k)] = \mathbb{P}[x_2(k) < \eta].$$

Therefore, the idea is to state the identification of $\eta$ in terms of the following optimization problem:

$$\min_{\eta} J'(\eta), \quad J'(\eta) \equiv \sum_{\ell=0}^{n} (\mathbb{E}[S_3(T_\ell)] - \hat{\mathbb{P}}[x_2(T_\ell) < \eta])^2,$$

where $\mathbb{E}[S_3(T_\ell)]$ and $\hat{\mathbb{P}}[x_2(T_\ell) < \eta]$ are suitable estimates of $\mathbb{E}[S_3(T_\ell)]$ and $\mathbb{P}[x_2(T_\ell) < \eta]$ drawn from multiple experiments on $\mu$ and $\mathcal{M}$, respectively. A general way to compute them is the following. Consider $H$ experiments on $\mathcal{M}$ with equal observation times $T_\ell$, with $\ell = 0, 1, \ldots, n$, and $h$ experiments on $\mu$. Let superscript "($i$)" denote outcomes of the $i$-th experiment (on $\mathcal{M}$ or on $\mu$). For $\ell = 0, 1, \ldots, n$ one computes:

$$\hat{\mathbb{E}}[S_3(T_\ell)] \equiv \frac{1}{h} \sum_{i=1}^{h} S_3^{(i)}(T_\ell) \simeq \mathbb{E}[S_3(T_\ell)],$$

$$\hat{\mathbb{P}}[x_2(T_\ell) < \eta] \equiv \frac{1}{H} \sum_{i=1}^{H} \mathbb{1}_\eta(y_2^{(i)}(T_\ell)) \simeq \mathbb{P}[x_2(T_\ell) < \eta],$$

where $\mathbb{1}_\eta$ is an indicator function that is equal to 1 if $y_2 < \eta$, and is zero otherwise. Note that this requires virtually no assumption on the observation times of $\mu$. Indeed, the method of Section III-A provides estimates of $S_3$ at the natural time scale $N$ regardless of the choice of $\tau$. The estimation method is expected to converge for $H, h \to \infty$. Yet, for small values of $H$ and $h$, good estimates may be obtained, provided $n$ is large enough.

VI. RESULTS

We tested the whole identification procedure on synthetic data generated by the $B.subtilis$ model with the true parameter values of Table I. Equally spaced observations of were collected every 1.2, 6 and 12 minutes over the time span
TABLE II

<table>
<thead>
<tr>
<th>$T_{\text{obs}}$ (min)</th>
<th>1.2</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$l_3$ (0.2)</td>
<td>0.1991</td>
<td>0.2088</td>
</tr>
<tr>
<td>$k_3$ (0.5)</td>
<td>0.4977</td>
<td>0.5214</td>
</tr>
<tr>
<td>$l_4$ (0.2)</td>
<td>0.1900</td>
<td>0.1829</td>
</tr>
<tr>
<td>$k_4$ (1)</td>
<td>0.9889</td>
<td>0.9449</td>
</tr>
<tr>
<td>$c_4$ (0.4)</td>
<td>0.3685</td>
<td>0.3747</td>
</tr>
<tr>
<td>$l_5$ (0.2)</td>
<td>0.2078</td>
<td>0.2112</td>
</tr>
<tr>
<td>$k_5$ (1)</td>
<td>1.0061</td>
<td>0.9922</td>
</tr>
<tr>
<td>$c_5$ (0.4)</td>
<td>0.4256</td>
<td>0.4479</td>
</tr>
</tbody>
</table>

ESTIMATES OF $\theta_\mu$ IN DIFFERENT EXPERIMENTAL CONDITIONS. $T_{\text{obs}}$ DENOTES THE TIME BETWEEN CONSECUTIVE MEASUREMENTS.

Fig. 3. Comparison between concentrations $x_3$, $x_4$, $x_5$ (solid line) and their estimated values at observation instants (dots).

$[0, 1200]$ min. These situations correspond to constant values of $\delta_i$ equal to 100, 500 and 1000, respectively, for a total of $N = 1000, 200$ and 100 observations. In what follows $T_{\text{obs}}$ will be used to denote the time between consecutive observations. Measurement noise levels were set to $\sigma_1 = 0.03$, $\sigma_2 = 0.12$, $\sigma_3 = \sigma_4 = \sigma_5 = 0.06$, yielding outcomes of $e_i$ within 10% of values of $x_i$ in the asymptotic regime.

A. Results for $\mu$

Table II shows the identification results obtained for $T_{\text{obs}}$ equals to 1.2 and 6 min. For $T_{\text{obs}} = 12$ min, as expected, observations were too sparse to capture the dynamical behavior of the three concentrations and therefore estimation could not be performed. Notice how all estimated parameters differ from their true value for less than 10%, except for $c_5$, whose relative estimation error is about 12% in the worst case. Figure 3 shows the noiseless behavior of $x_3$, $x_4$, $x_5$ and the corresponding estimates $\hat{x}_3$, $\hat{E}[x_4]$, $\hat{E}[x_5]$ at observation instants with $T_{\text{obs}} = 6$ min, highlighting the good fit of the estimation results.

B. Results for $\mathcal{M}$

For applying the PEM identification described in Section IV, mean and variance of $x_5$ were computed empirically at times $\tau_2 = \ell \times 12$ min, with $\ell = 1, \ldots, 100$, from 50 independent experiments on $\mu$. Their value over the natural time scale $\mathbb{N}$ was then estimated by spline interpolation. In order to make the optimization of $J_{1,2}$ more robust with respect to local minima, numerical minimization was started from 5 different parameter guesses spread over two orders of magnitude about the true parameter values. Table III shows the identification results obtained for the three different observation rates, along with the minimum value found for $J_{1,2}$. There is an excellent agreement between the estimated values of $r$, $X_0$, $D_{\text{max}}$, $k_1$, $k_2$ and their true values, reported in parenthesis in the first column of Table III. The increasing values of minimum $J_{1,2}$ indicate a small loss of performance for larger values of the observation period. Figure 4 shows the extended Kalman state predictions corresponding to the true parameters and those estimated with $T_{\text{obs}} = 12$ min. It can be noticed that predictions are very similar. However, predictions with the estimated parameters tend to oversmooth the state trajectory. Note how predictions tend to settle to a constant value, although real state values oscillate. In this regime, predictions with estimated parameters also show a small bias with respect to the optimal ones.

C. Results for $\eta$

Estimates of $\eta$ for changing number of experiments and observation period are reported in Table IV. They were all obtained by standard numerical minimization with initial guesses 1, 5 and 10. The numbers of experiments $h$ and $H$ were taken to be the same in each case. Convergence to reasonable estimates was verified in all cases. In general, it is confirmed that accuracy grows with the number of samples and with the number of experiments. However, improvements are limited above small values of $H$, which suggests that our method is suitable for use with few real-world experiments.
We have studied parameter identification for the stochastic hybrid model of \textit{B. subtilis} subtilin production proposed in [16]. We assumed that data result from a sparse and possibly irregular sampling of the system state. We considered measurements of the various components of the state separately, in accordance with their different biological nature. This setting matches the constraints of current experimental techniques in molecular biology.

We reformulated the model as the composition of macroscopic and microscopic dynamical systems coupled by two static systems in a feedback structure. This allowed us to derive methods for parameter identification at both population-level (growth rate, nutrient consumption rate etc.) and cell-level (protein synthesis and decay rates, probabilities of gene expression) based on measurements at the same biological scale. To do this, we borrowed techniques from deterministic hybrid systems and introduced new ideas for the estimation of stochastic switching dynamics. We tailored our methods on a typical structure of genetic regulatory chains.

This research may be extended in several ways. One is the validation on real-world data, which may lead to a refinement of the model and to an optimized design of the biochemical experiments. Another direction is to reformulate \textit{B. subtilis} identification based on the continuous-time piecewise-deterministic Markov model of [19]. In addition, it is our intention to extend this work to more general genetic regulatory mechanisms. Finally, theoretical analysis of the performance of the estimators is a long-term aim.

REFERENCES


