Local identification of piecewise deterministic models of genetic networks

Eugenio Cinquemani, Andreas Milias-Argeitis, Sean Summers, and John Lygeros

Institut für Automatik, ETH Zurich, Switzerland
{cinquemani,milias,summers,lygeros}@control.ee.ethz.ch

Abstract. We address the identification of genetic networks under stationary conditions. A stochastic hybrid description of the genetic interactions is considered and an approximation of it in stationary conditions is derived. Contrary to traditional structure identification methods based on fitting deterministic models to several perturbed equilibria of the system, we set up an identification strategy which exploits randomness as an inherent perturbation of the system. Estimation of the dynamics of the system from sampled data under stability constraints is then formulated as a convex optimization problem. Numerical results are shown on an artificial genetic network model. While our methods are conceived for the identification of interaction networks, they can as well be applied in the study of general piecewise deterministic systems with randomly switching inputs.

Key words: Piecewise deterministic systems, state-space identification, Markov processes, sampled systems, convex optimization.

1 Introduction

Genetic regulatory networks govern the synthesis of proteins in the living cell, and are thus responsible for fundamental cell functions such as metabolism, development and replication. 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 [1]. 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 ([2–6], among others) who recently applied hybrid systems tools in this context. In addition, mounting experimental evidence suggests that gene expression, both in prokaryotes and eukaryotes, is an inherently stochastic process. Stochasticity can be attributed to the randomness of the transcription and translation processes (intrinsic noise), as well as to fluctuations in the amounts of molecular components that affect the expression of a certain gene (extrinsic noise), see [7–10]. In [11], stochastic modelling of genetic regulatory networks is reviewed along with numerical simulation methods and is

compared to deterministic modelling. The authors have addressed stochastic hybrid modelling of genetic networks in [12]. A similar approach is taken in [13] for the analysis and numerical simulation of basic transcriptional network modules.

Recent works — [4, 14–17] — have started to address the problem of learning genetic network models from experimental data. In particular, the literature on identification of stochastic regulatory network models is quite new [18–20, 12]. A central problem in genetic network modelling is the identification of the network of interactions. Traditional approaches based on dynamic modelling rely on matching deterministic models to different equilibria corresponding to known perturbations of the system, see e.g. [21–23]. That is, one assumes that protein concentrations x evolve according to a kinetic model $\dot{x} = f(x, u)$, where u is a known perturbation input acting on the system. Then, a linearized model $\dot{x} = Ax + Bu$ is identified around several equilibrium points of the system corresponding to different constant values of u. This turns identification into a regression problem $0 = AX + BU$, where X is a matrix of observed equilibria and U is the matrix of the corresponding inputs. Matrix A carries information about the structure of the interaction network, hence the interest in its estimation. The main drawback of this approach is due to the assumption that A is the same at all equilibria. This implies that perturbations must be small. At the same time, several equilibria must be explored for the solution of the regression to be unique. The inherent random perturbations of the dynamics are not exploited in this case, in that the choice of deterministic modelling simply ignores this contribution.

In this paper we address identification of the structure of the network in a stochastic hybrid modelling framework. We start from the model described in [12] and consider a stochastic approximation of it around a stationary point of the system. Based on this, we borrow tools from the theory of identification of linear stochastic processes [24] to estimate the structure of the system. The conceptual difference with respect to traditional methods is that we make use of the randomness driving the system as a natural perturbation of the dynamics, with no further assumptions on the invariance of the dynamics. Artificial perturbations corresponding to several stationary conditions may be used to improve the estimation results and to separate different contributions, e.g. spontaneous degradation from regulatory effects. The identification procedure we propose relies on a local approximation of the stochastic hybrid model with a continuous stochastic model. This simplifies the identification problem but certain details of the network structure are lost in the approximation. The identification methods presented in [17, 12], which build on the stochastic hybrid structure of the system, may then be used to recover the model in full detail.

The contribution of the paper is twofold. First, we introduce an approach to genetic network structure identification that accounts for and takes advantage of the inherent stochasticity of the systems. Of course, the approach requires that this randomness be reflected in the data. In view of the rapid progress of the protein level measurement techniques and of the advent of single-cell experiments [25, 26], we believe that this approach is going to be applicable to

experimental data in the near future. Second, on a more theoretical level, we provide methods for the approximation and the identification of a family of stochastic hybrid models (namely the class of piecewise deterministic processes with switching inputs) that is relevant to a number of application scenarios.

The paper is organized as follows. In Section 2, we describe our stochastic hybrid framework for genetic network modelling. An approximate model of the stochastic hybrid dynamics under stationary conditions is derived in Section 3. Section 4 states the structure identification problem of our concern and describes a solution based on convex numerical optimization. A discussion of the method and of its possible extensions is developed in Section 5. The performance of our method is discussed in Section 6 by way of numerical experiments. Conclusions on and perspective of our work are drawn in Section 7. Mathematical proofs are included in the appendix.

2 Piecewise deterministic models of genetic networks

A genetic network may be thought of as a collection of n proteins and of n corresponding genes along with their regulatory interactions. New molecules of a protein are synthesized when the gene that encodes it is expressed. The expression of a gene is regulated by one or more transcription factors (TFs). These are themselves proteins encoded by the genes of the network. In the simplest case, if a transcription factor is an activator (inhibitor), its binding to the promoter of the gene will activate (inhibit) a cascade of reactions that ultimately leads to the synthesis of new molecules of the protein encoded by that gene. In more generality, the simultaneous presence/absence of several transcription factors at the promoter site determines the status of the gene expression. We assume that changes in protein concentration due to synthesis and spontaneous degradation are well approximated by deterministic (kinetic) equations. On the other hand, the inherent randomness driving the binding/unbinding events and the presence of a limited number of binding sites leads us to model initiation and termination of gene expression as a stochastic process.

For a fixed $T \in \mathbb{R}_+$, let $\mathscr{T} = T \cdot \mathbb{N} = \{T, 2T, 3T, \ldots\}$ be a sequence of time instants. For $t \in \mathscr{T}$, let $x(t) \in \mathbb{R}_+^n$ be a continuous state vector of protein concentrations. Let $\ell(i) \subset \{1, \ldots, n\}$ denote the set of proteins acting as TFs on gene i. For each $k \in \ell(i)$ and $t \in \mathscr{T}$, let $u_{i,k}(t) \in \{0, 1\}$ be a discrete state variable that encodes the presence ($u_{i,k}(t) = 1$) or absence ($u_{i,k}(t) = 0$) of TF k at the promoter site of gene i. Therefore the activity of gene i is governed by a discrete state taking values in $\{0, 1\}^{|\ell(i)|}$, where $|\cdot|$ denotes set cardinality. Let $u(t) \in \{0, 1\}^m$, with $m = |\ell(1)| + \ldots + |\ell(n)|$, be a vector collecting all discrete variables $u_{i,k}(t)$, with $i = 1, \ldots, n$ and $k \in \ell(i)$. We model the evolution in time of the protein concentrations due to regulated synthesis and spontaneous degradation by the discrete-time dynamical equation

$$x(t + T) = Ax(t) + g\big(u(t + T)\big), \tag{1}$$

where $A \in \mathbb{R}_+^{n \times n}$ is a diagonal matrix of spontaneous degradation rates and $g : \mathbb{R}^m \to \mathbb{R}_+^n$ is a smooth function that quantifies the rate of synthesis of new

proteins in terms of the discrete state u. Typically, each component g_i of g takes the form

$$g_i(u) = \sum_j b_i^j \prod_{k \in \ell(i,j)} u_{i,k}, \tag{2}$$

where $b_i^j \in \mathbb{R}$ and the $\ell(i,j) \subseteq \{1,\dots,n\}$ are such that $\cup_j \ell(i,j) = \ell(i)$. To fix the ideas, each term of the summation corresponds to a different gene activation path, and b_i^j is the corresponding synthesis rate for protein i.

Stochasticity comes in the model by the description of the binding events, i.e. of the discrete transitions of u. Let $(\Omega, \mathscr{E}, \mathbb{P})$ be a probability space. For $t \in \mathscr{T}$, we describe the transitions of every $u_{i,k}$ as discrete random events with probabilities $\mathbb{P}[u_{i,k}(t+T)|u_{i,k}(t), x_k(t)]$ depending on the current protein concentrations $x_k(t)$ (e.g. the larger the concentration x_k, the larger the probability that a molecule of protein k binds to the promoter site of gene i). In light of this and Eq. (1), $u : \mathscr{T} \times \Omega \to \{0,1\}^m$ and $x : \mathscr{T} \times \Omega \to \mathbb{R}_+^n$ are two random processes defined on $(\Omega, \mathscr{E}, \mathbb{P})$. For simplicity we shall keep writing $x(t)$ and $u(t)$ in place of $x(t,\omega)$ and $u(t,\omega)$, where $\omega \in \Omega$. We impose the following two assumptions:

Assumption 1 *For every $t \in \mathscr{T}$, $u(t+T)$ and $x(t+T)$ are conditionally independent from the past history $x^-(t) = \{x(0), x(T), \dots, x(t-T)\}$ and $u^-(t) = \{u(0), u(T), \dots, u(t-T)\}$ given $x(t)$ and $u(t)$.*

Assumption 2 *For all $t \in \mathscr{T}$, the transition probability law*

$$p_{v,v'}(z) = \mathbb{P}[u(t+T) = v' | u(t) = v, x(t) = z], \qquad v, v' \in \{0,1\}^m,\ z \in \mathbb{R}_+^n, \tag{3}$$

is independent of t.

For a fixed initial condition $x(0) = x_0$ and an initial probability distribution $p_v^0 = \mathbb{P}[u(0) = v]$, the above completely specifies the probability laws of the joint process (x,u). A straightforward consequence of Assumptions 1 and 2 is

Proposition 1. *The joint process $\big(x(t), u(t)\big)$ is Markovian. For $t \in \mathscr{T}$,*

$$\begin{aligned}
&\mathbb{P}[x(t+T) = z', u(t+T) = v' | x(t) = z, u(t) = v, x^-(t), u^-(t)] = \\
&= \mathbb{P}[x(t+T) = z' | x(t) = z, u(t+T) = v'] \mathbb{P}[u(t+T) = v' | x(t) = z, u(t) = v] \\
&= \begin{cases} 0, & \text{if } z' \neq Az + g(v'), \\ p_{v,v'}(z), & \text{otherwise.} \end{cases}
\end{aligned}$$

For a given value of $u(t+T)$, Eq. (1) describes the transition of the continuous-valued process x from $x(t)$ to $x(t+T)$. We call $x(t)$ a piecewise deterministic process in that, as long as the value of u remains unchanged, the evolution of x is deterministic. On the other hand, for fixed values of $x(t)$ and $u(t)$, Eq. (3) determines the random outcome of the discrete-valued process $u(t+T)$. The joint process (x,u) resulting from the interconnection of the two processes is thus stochastic and hybrid. It is easy to recognize the class of processes defined above as a discrete-time variant of the Piecewise Deterministic Markov Processes introduced by [27].

To keep the analysis tractable we shall make a further assumption.

Assumption 3 *For every $t \in \mathscr{T}$, $u(t+T)$ is independent of $u(t)$ given $x(t)$.*

Biologically relevant conditions under which this assumption holds are discussed in [12, 17]. Since $p_{v,v'}(z)$ is independent of v, we shall replace $p_{v,v'}(z)$ by $p_{v'}(z) = \mathbb{P}[u(t+T) = v'|x(t) = z]$.

3 Stochastic approximation under stationary conditions

The stochastic hybrid structure of the genetic network model makes analysis and identification quite challenging. In [17] and [12] we proposed global methods for the identification of unknown model parameters that build on the stochastic hybrid model structure. The identification results are very good in that context since the full knowledge of the system structure was exploited. Yet, the associated optimization problems are nonconvex and generally hard to solve. Here we address the more difficult problem of structure identification and take an alternative approach to solve it. We approximate the stochastic hybrid dynamics locally by a continuous stochastic system and match the latter to the data. The resulting optimization problem is tractable, but the structure of the original stochastic hybrid model is partly obscured. In principle, the methods presented in [17, 12] allow one to re-introduce the details of the network structure. This may be achieved via a series of heuristics which are currently being developed.

Assume that the joint process $\big(x(t), u(t)\big)$ has reached stationarity. Define $\bar{x} = \lim_{t\to\infty} \mathbb{E}[x(t)]$ and $\bar{u} = \lim_{t\to\infty} \mathbb{E}[u(t)]$, where $\mathbb{E}[\cdot]$ denotes expectation. Using the first-order expansion

$$g(u) = g\big(\bar{u}\big) + G_{\bar{u}}(u - \bar{u}) + o(u - \bar{u}) \simeq g(\bar{u}) + G_{\bar{u}} \cdot (u - \bar{u}),$$

where $G_{\bar{u}}$ is the Jacobian of g evaluated at $\bar{u}$, one may write

$$x(t+T) = Ax(t) + g\big(u(t+T)\big) \simeq Ax(t) + g(\bar{u}) + G_{\bar{u}} \cdot \big(u(t+T) - \bar{u}\big), \qquad (4)$$

the approximation being most accurate for small variance of $g\big(u(t)\big)$. Define $\tilde{x}(t) = x(t) - \bar{x}$ and $\tilde{u}(t) = u(t) - \bar{u}$.

Proposition 2. *Assume that (4) holds as an equality. Then $\bar{x} = A\bar{x} + g(\bar{u})$ and*

$$\tilde{x}(t+T) = A\tilde{x}(t) + G_{\bar{u}}\tilde{u}(t+T). \qquad (5)$$

We shall call (5) the approximate linear model for $\tilde{x}$. Of course, the model is not truly linear due to the dependence of $\tilde{u}$ on $\tilde{x}$. For any $\tilde{v} = v - \bar{u}$, with $v \in \{0,1\}^m$, and $\tilde{z} = z - \bar{x}$, with $z \in \mathbb{R}^n_+$, define $\tilde{p}_{\tilde{v}}(\tilde{z}) = \mathbb{P}[\tilde{u}(t+T) = \tilde{v}|\tilde{x}(t) = \tilde{z}]$.

Proposition 3. $\tilde{p}_{\tilde{v}}(\tilde{z}) = p_v(z)$.

Along with Eq. (5), this straightforward result provides locally an approximate model for the stochastic hybrid process $\big(\tilde{x}, \tilde{u}\big)$.

We are interested in the (approximate) second-order moments of the piecewise deterministic process $\tilde{x}(t)$. By definition, $\mathbb{E}[\tilde{x}(t)] = 0$. For $\ell \in \mathbb{Z}$, define the covariance function $\Sigma_x(\ell) = \mathbb{E}[\tilde{x}(t+\ell T)\tilde{x}(t)^T]$. By stationarity $\Sigma_x(\ell) = \Sigma_x(-\ell)^T$, and we may restrict our attention to $\ell \in \mathbb{N}$. Note that $\Sigma_x(0)$ is the covariance matrix of $\tilde{x}$.

Assumption 4 *There exists $F_{\bar{x}} \in \mathbb{R}^{n\times n}$ such that, for all $t \in \mathbb{Z}$,*

$$\mathbb{E}[\tilde{u}(t+T)|x(t)] = F_{\bar{x}}\tilde{x}(t). \tag{6}$$

Proposition 4. *For every $\ell \in \mathbb{N}$ it holds that*

$$\Sigma_x(\ell+1) = (A + G_{\bar{u}}F_{\bar{x}})\Sigma_x(\ell), \tag{7}$$

where $\Sigma_x(0)$ is the solution of

$$\Sigma_x(0) = (A + G_{\bar{u}}F_{\bar{x}})\Sigma_x(0)(A + G_{\bar{u}}F_{\bar{x}})^T + G_{\bar{u}}QG_{\bar{u}}^T. \tag{8}$$

In turn, $Q = \mathbb{E}\big[\mathrm{Var}\big(u(t+T)|x(t)\big)\big]$, where $\mathrm{Var}(\cdot|\cdot)$ denotes conditional variance.

Assumption 4 is met if $f(x) = \mathbb{E}[u(t+T)|x(t) = x]$ is linear. In practice, we shall assume that this is a valid approximation in a neighborhood of $\bar{x}$, i.e.

$$f(x) = f(\bar{x}) + F_{\bar{x}}\tilde{x} + o(\tilde{x}) \simeq f(\bar{x}) + F_{\bar{x}}\tilde{x}.$$

In this case, $\bar{u} = \mathbb{E}[u(t+T)] = \mathbb{E}\big[\mathbb{E}[u(t+T)|x(t)]\big] \simeq \mathbb{E}[f(\bar{x}) + F_{\bar{x}}\tilde{x}(t)] = f(\bar{x})$. Therefore, Assumption 4 is just a consequence of

$$\mathbb{E}[\tilde{u}(t+T)|x(t)] = \mathbb{E}[u(t+T)|x(t)] - \bar{u} \simeq \big(f(\bar{x}) + F_{\bar{x}}\tilde{x}(t)\big) - f(\bar{x}) = F_{\bar{x}}\tilde{x}(t).$$

Proposition 4 implies that the approximate second-order moments of the piecewise deterministic process $\tilde{x}$ are equal to those of a process described by the linear stationary state-space model

$$\tilde{x}(t+T) = \mathbb{A}_{\bar{x},\bar{u}}\tilde{x}(t) + G_{\bar{u}}w(t), \tag{9}$$

where $\mathbb{A}_{\bar{x},\bar{u}} = A + G_{\bar{u}}F_{\bar{x}}$ and $w(\cdot)$ is an i.i.d. process uncorrelated with $x^-(t)$ with mean zero and covariance matrix Q. Interestingly, this corresponds to replacing $\tilde{u}(t+T)$ in (5) with

$$\tilde{u}(t+T) = F_{\bar{x}}\tilde{x}(t) + w(t), \tag{10}$$

i.e. the gene regulation encoded by $\tilde{u}(t+T)$ may locally be thought of as a static linear state feedback with matrix gain $F_{\bar{x}}$ and additive noise w.

4 Constrained identification of the linearized model

The approximation of the second-order moments of $\tilde{x}$ with those of (9) allows us to use concepts from the theory of linear stationary processes for the analysis and identification of piecewise deterministic systems. In view of the application to genetic network modelling, we are primarily interested in the estimation of the matrix $\mathbb{A}_{\bar{x},\bar{u}}$. This matrix combines spontaneous protein degradation (diagonal matrix A) with the effects of the regulatory interactions (matrix $G_{\bar{u}}F_{\bar{x}}$). In particular, $G_{\bar{u}}$ reflects the topology of the network, whereas $F_{\bar{x}}$ reflects the probability of each individual regulatory event near the stationary point $(\bar{x}, \bar{u})$. Note that the off-diagonal elements of $\mathbb{A}_{\bar{x},\bar{u}}$ only depend on the product $G_{\bar{u}}F_{\bar{x}}$.

As a result, the sign of each element $\left[\mathbb{A}_{\bar{x},\bar{u}}\right]_{i,k}$, with $i \neq k$, reveals the average (positive or negative) regulatory effect of protein k on the expression of gene i. A zero element, on the other hand, suggests that protein k is not involved in the regulation of gene i, at least around the stationary point $(\bar{x}, \bar{u})$. Therefore, the identification of $\mathbb{A}_{\bar{x},\bar{u}}$ provides information on the structure of the regulation network. A priori knowledge on the system (existing or non-existing interactions, for instance) should be accounted for at this stage. In this section we shall only constrain matrix $\mathbb{A}_{\bar{x},\bar{u}}$ to be stable. Local stability is a fundamental property of genetic regulatory networks near equilibria and is also central for Approximation 5. If $\mathbb{A}$ were unstable, process (10), that is (5), would not be second-order stationary. From now on, we assume that $(\tilde{x}, \tilde{u})$ satisfies (5) and (6).

Assume that measurements y of (the protein concentrations) x are collected every $N > 0$ samples. This is captured by the following model:

$$y(\tau) = x(\tau) + n(\tau), \qquad \tau \in NT \times \mathbb{Z},$$

where n is a white noise process (not necessarily Gaussian), uncorrelated with x, with mean zero and covariance matrix $R = \mathbb{E}[nn^T]$. The identification problem is formulated as follows.

Problem 1. Given $M+1$ data points $\mathscr{Y} = \big\{y(t), y(t+NT), \ldots, y\big(t+MNT\big)\big\}$, compute an estimate $\widehat{\mathbb{A}}$ of $\mathbb{A}$ such that $\widehat{\mathbb{A}}$ is stable.

The case where $N > 1$ is especially relevant to genetic network identification. In this context, the discrete network events occur at a time scale T which is usually smaller than the time period that separates subsequent experimental measurements. Let $\bar{y} = \mathbb{E}[y(t)] = \mathbb{E}[x(t)] = \bar{x}$ be the mean of y and let $\tilde{y} = y - \bar{y}$. In practice, $\bar{y}$ can be estimated and removed from the data. For $\ell \in \mathbb{Z}$, define the covariance function $\mathit{\Lambda}(\ell N) \triangleq \mathbb{E}[\tilde{y}(t+\ell NT)\tilde{y}(t)^T]$. Note that $\mathit{\Lambda}(-\ell N) = \mathit{\Lambda}(\ell N)^T$ and that $\mathit{\Lambda}(0)$ is the covariance matrix of y.

Proposition 5. *$\mathit{\Lambda}(N) = \mathbb{A}^N\big(\mathit{\Lambda}(0) - R\big)$ and, for $\ell > 0$, $\mathit{\Lambda}(\ell N + N) = \mathbb{A}^N \mathit{\Lambda}(\ell N)$.*

For $\ell = 0, 1, \ldots, L$ with $L << M$, one may compute empirical estimates $\widehat{\mathit{\Lambda}}(\ell N)$ of $\mathit{\Lambda}(\ell N)$ as follows:

$$\widehat{\mathit{\Lambda}}(\ell N) = \frac{1}{M-\ell} \sum_{h=0}^{M-\ell} \tilde{y}(t+\ell NT+hNT)\tilde{y}(t+hNT)^T.$$

The approximation $\widehat{\mathit{\Lambda}}(\ell N) \simeq \mathit{\Lambda}(\ell N)$ is most accurate as $M \to \infty$. Assume for the time being that R is known. Define the $\mathbb{R}^{n \times nL}$ matrices

$$\widehat{\mathit{\Lambda}}_+ = \Big[\widehat{\mathit{\Lambda}}(N)\ \widehat{\mathit{\Lambda}}(2N) \cdots\ \widehat{\mathit{\Lambda}}(LN)\Big], \quad \widehat{\mathit{\Lambda}}_- = \Big[\big(\widehat{\mathit{\Lambda}}(0) - R\big)\ \widehat{\mathit{\Lambda}}(N) \cdots\ \widehat{\mathit{\Lambda}}\big((L-1)N\big)\Big].$$

In the light of Proposition 5, we address the identification Problem 1 by seeking a solution to the following optimization problem in the unknown matrix $\mathbb{A}$:

$$\text{minimize} \quad ||\widehat{\mathit{\Lambda}}_+ - \mathbb{A}^N \widehat{\mathit{\Lambda}}_-|| \qquad \text{subject to} \quad \mathbb{A} \text{ stable},$$

where $||\cdot||$ denotes the matrix spectral norm. In general, this problem is noncovex due to matrix exponentiation. To circumvent this issue, we use the fact that $\mathbb{A}$ is stable if and only if $\mathbb{A}^N$ is stable. We propose to solve the problem in two steps:

1. minimize $||\widehat{\Lambda}_+ - X\widehat{\Lambda}_-||$ subject to X stable. Denote the solution by $\widehat{X}$;
2. compute the matrix N-th root $\widehat{X}^{1/N}$.

Step 1. This amounts to matching the matrix $X = \mathbb{A}^N$ to the available covariance data in accordance with Proposition 5. By the Lyapunov theorem, the stability constraint on X is equivalent to the existence of a positive definite matrix P such that $XPX^T - P < 0$. Using Schur complement, this can be turned into the equivalent LMI

$$\begin{bmatrix} P & AP \\ PA^T & P \end{bmatrix} > 0,$$

with unknowns A and P. Define $Z = PA$. Using a series of standard transformations [28, 29] based on the properties of the spectral norm, the problem can be reformulated in terms of the convex optimization

$$\text{minimize} \quad ||P\widehat{\Lambda}_+ - Z\widehat{\Lambda}_-|| \qquad \text{subject to} \quad \begin{bmatrix} P - \epsilon I & Z^T \\ Z & P \end{bmatrix} \geq 0, \quad P \geq I,$$

with unknowns Z and P. Here $\epsilon \in \mathbb{R}_+$ is a small design constant used to make the constraint set closed and to ensure the strict stability of the solution. This problem has a unique solution whenever the matrix $\widehat{\Lambda}_-$ has full row rank. Denote the solution with $\hat{Z}$ and $\hat{P}$. Then, setting $\widehat{X} = \widehat{P}^{-1}\widehat{Z}$ provides an approximate solution to the original problem.

Step 2. This requires the computation of the N-th root of a square matrix. The choice of the N-th matrix root is nonunique, see [30] for a detailed characterization of the solutions. However, provided the sampling time T is small enough, we expect that $\mathbb{A}$ is close to the identity, that is, all its eigenvalues should be located in a neighborhood of 1. Based on this consideration, we choose to compute the principal N-th root. By definition, this is the unique root matrix having all eigenvalues λ such that $\arg(\lambda) \in [-\pi/2N, \pi/2N]$, i.e. having all eigenvalues in the sector of the complex plain containing 1. Several algorithms for computing the principal root exist [30, 31].

If R is unknown, we modify the problem by removing the leftmost $\mathbb{R}^{n\times n}$ element from $\widehat{\Lambda}_+$ and $\widehat{\Lambda}_-$. That is, we define the $\mathbb{R}^{n\times(L-1)n}$ matrices

$$\widehat{\Lambda}_+ = \left[\widehat{\Lambda}(2N)\ \widehat{\Lambda}(3N) \cdots\ \widehat{\Lambda}(LN)\right], \quad \widehat{\Lambda}_- = \left[\big(\widehat{\Lambda}(N)\ \widehat{\Lambda}(2N) \cdots\ \widehat{\Lambda}\big((L-1)N\big)\right]$$

and perform Steps 1 and 2 with these new matrices to get the estimate $\widehat{X}$. Provided $\widehat{X}$ is invertible, an estimate $\widehat{R}$ of R may be computed by solving $\widehat{\Lambda}(N) = \widehat{X}(\widehat{\Lambda}(0) - \widehat{R})$.

5 Discussion and extensions

We mentioned above that the stability constraint can be equally imposed on $\mathbb{A}$ or on $\mathbb{A}^N$. To simplify the identification procedure, we decided to enforce this constraint on matrix $\mathbb{A}^N$ in the first identification step. In general, different information on the matrix $\mathbb{A}$, i.e. the sign of certain elements or the sparsity of the matrix, does not carry over to the matrix $\mathbb{A}^N$. If such prior knowledge on $\mathbb{A}$ is available, it is convenient to turn the identification scheme into a three-step procedure:

1'. minimize $||\widehat{\Lambda}_+ - X\widehat{\Lambda}_-||$ with respect to X, and name the solution $\widehat{X}$;
2'. compute the matrix N-th root $\widehat{X}^{1/N}$;
3'. minimize $||\widehat{X}^{1/N} - \mathbb{A}||$ subject to (stability and other) constraints on $\mathbb{A}$.

Step $1'$ is an easy convex problem and serves the purpose of matching $X = \mathbb{A}^N$ to the data without constraints. If the spectral norm is replaced by the Frobenius norm, then the solution can be computed explicitly as $\widehat{X} = \widehat{\Lambda}_+\widehat{\Lambda}_-^R$, where $\widehat{\Lambda}_-^R = \widehat{\Lambda}_-^T\big(\widehat{\Lambda}_-\widehat{\Lambda}_-^T\big)^{-1}$ is the Moore-Pennrose pseudo-inverse of $\widehat{\Lambda}_-$. Step $2'$ is the same as the former step 2 but yields an unconstrained root matrix $\widehat{X}^{1/N}$. Finally, step 3 seeks a constrained approximation of $\widehat{X}^{1/N}$ using the prior information on $\widehat{\mathbb{A}}$. Effective heuristics to solve this problem by convex optimization exist for many constraints of interest, see e.g. [21], and will not be further discussed here.

Given an estimate $\widehat{\mathbb{A}}_{\bar{x},\bar{u}}$, one cannot separate (the diagonal matrix) A from (the diagonal elements of) $G_{\bar{u}}F_{\bar{x}}$ and, in turn, $G_{\bar{u}}$ from $F_{\bar{x}}$. This is the same limitation of traditional methods. In these methods, however, perturbations of the system such as gene enhancement or knock-out are used to infer the overall system dynamics. In our setting, the overall dynamics $\mathbb{A}_{\bar{x},\bar{u}}$ are estimated based on a fixed experimental scenario, while system perturbations (i.e. estimates corresponding to different stationary points $(\bar{x},\bar{u})$) may be exploited to discern the individual contributions of A, $G_{\bar{u}}$ and $F_{\bar{x}}$.

In addition to the estimation of matrix $\mathbb{A}$, our local approximation of the stochastic hybrid model can be used to learn the dimension of the system from the data. Consider for simplicity $N = 1$. It is well known that the rank of the block Hankel matrix

$$H = \begin{bmatrix} \Lambda(1) & \Lambda(2) & \Lambda(3) & \cdots \\ \Lambda(2) & \Lambda(3) & \Lambda(4) & \cdots \\ \Lambda(3) & \Lambda(4) & \Lambda(5) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

associated with the linear system (9) is equal to the dimension of the state of a minimal realization of the system. Since the dimension of the state-space model (9) and of the piecewise deterministic model (1) are the same, the rank of H is indication of the dimension of the continuous state of (1). In the context of genetic network modelling, this type of analysis and other tools from the theory of realization/identification of linear stochastic processes [24] may help to develop methods for the estimation of the number of genes involved in the regulation of the observed proteins of the network.

6 Numerical experiments

We consider an interaction network with four genes. The system is described by the stochastic hybrid model

$$\begin{aligned} x_1^+ &= \lambda_1 x_1 + b_1 u_{1,1}^-(1 - u_{1,2}^+ u_{1,3}^-), & x_3^+ &= \lambda_3 x_3 + b_3^1 u_{3,1}^+ u_{3,2}^+ u_{3,3}^-, \\ x_2^+ &= \lambda_2 x_2 + b_2 u_{2,1}^-(1 - u_{2,2}^+ u_{2,3}^-), & x_4^+ &= \lambda_4 x_4 + b_4^1 u_{4,1}^+ + b_4^2. \end{aligned} \tag{11}$$

It is easy to verify that the protein synthesis rates are in the form (2). Processes $u_{i,k}^{\pm}(t+T)$ are independent given the current continuous state $x(t)$. The superscript $+$ or $-$ indicates whether the probability of $u_{i,k}(t+T)$ being equal to one is given by the sigmoidal function $\sigma_{i,k}^+(x_k) = x_k^d/(x_k^d + \theta^d)$ or by the complementary sigmoid $\sigma_{i,k}^-(x_k) = 1 - \sigma_{i,k}^+(x_k)$. Parameters $d \in \mathbb{R}_+$ and $\theta \in \mathbb{R}_+$ generally also depend on i, k. This model is in fact part of a larger model model for the nutrients stress response of bacterium *Escherichia Coli*. The interested reader is deferred to [12] and references therein for a more detailed discussion.

Using biologically plausible parameter values and initial conditions [12], it can be observed by simulation that system (11) eventually reaches a stationary regime. Sample trajectories from the stationary regime are plotted in Fig. 1. We

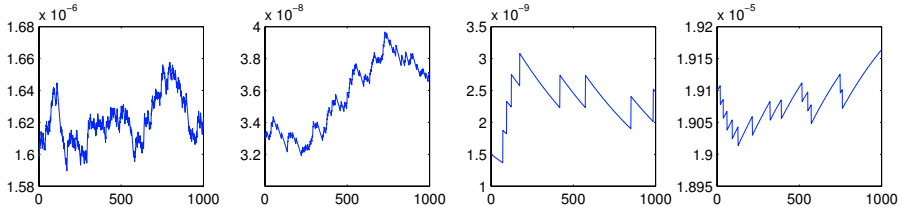

Fig. 1. Simulated trajectories of model (11) in stationary regime.

perform local identification of the above model from simulated data. Estimation of the matrix $\mathbb{A}_{\bar{x},\bar{x}}$ is performed on the basis of a single trajectory $y(t)$, with $1 \leq t \leq 1000$. We considered four different experimental scenarios:

A. No measurement noise, no undersampling ($N = 1$)
B. With measurement noise, no undersampling ($N = 1$)
C. No measurement noise, with undersampling ($N = 10$)
D. With measurement noise, with undersampling ($N = 10$)

When $N = 1$, all 1000 data points $y(1), y(2), \ldots, y(1000)$ are used. When $N = 10$, only the 100 data points $y(10), y(20), \ldots, y(1000)$ are used. This simulates two biological experiments of the same duration but with different sampling rates. Such small size of the data sets reflects the typical experimental practice where a limited number of protein concentration measurements are collected sparsely in time by a single biological experiment. Noise, when applicable, is

drawn from a normal distribution with mean zero and covariance matrix $R = \mathrm{diag}(r_1^2, r_2^2, r_3^2, r_4^2)$, with $r_i = 0.01 \cdot \bar{x}_i$. In the identification process, the stationary mean value $\bar{y}$ is computed empirically and removed from the data y. Then, the two-step identification procedure (with $L = 2$ and R known) is applied to the data. For each of the four scenarios above, 100 estimates of the matrix $\mathbb{A}_{\bar{x},\bar{u}}$ are drawn from 100 random simulations on the model. For comparison, the true value of $\mathbb{A}_{\bar{x},\bar{u}}$ is computed from Eqn. (11), where the mean values $\bar{x}$ and $\bar{u}$ are computed empirically from the simulated trajectories. The mean value and the variance of the estimates of all elements of $\mathbb{A}_{\bar{x},\bar{u}}$ are reported in Fig. 2 along with the true values. In all cases, estimates are affected by very little or no bias

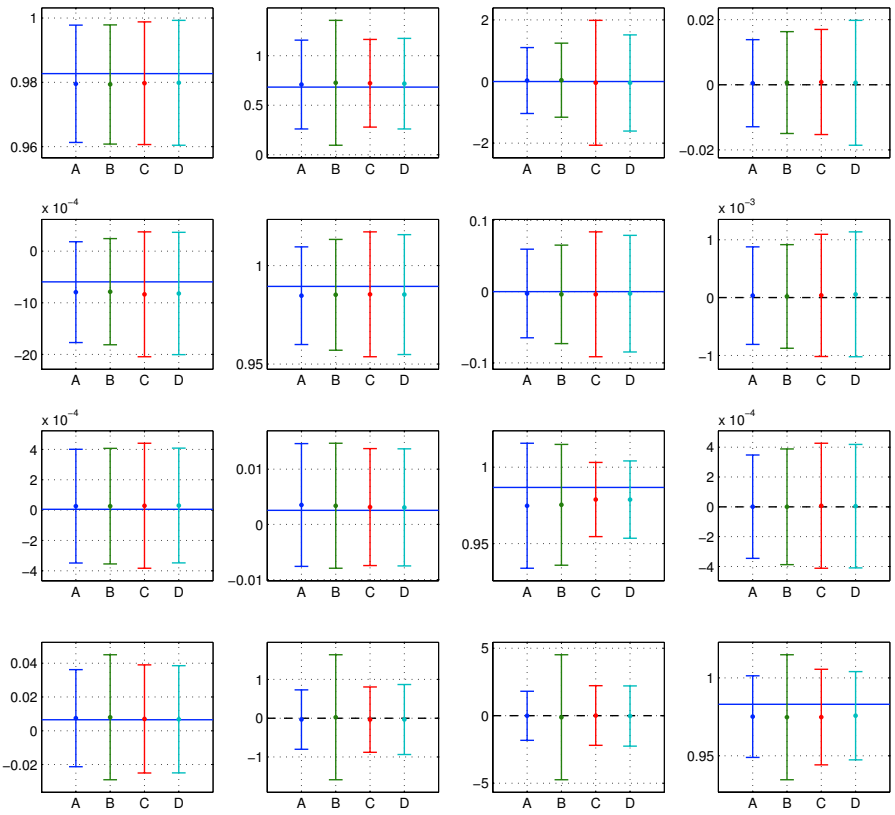

Fig. 2. For $r, c = 1, \ldots, 4$, the plot in the r-row and c-th column reports the estimation results for the (r, c)-th entry of $\mathbb{A}_{\bar{x},\bar{u}}$. In each plot, dots indicate the mean of the estimates and vertical bars correspond to 3 times the standard deviation of the estimates for the identification scenarios A (blue, left), B (green, second-left), C (red, second-right), D (cyan, right). Horizontal lines indicate the true entry values (dashed black lines for zeros, solid blue lines otherwise).

which appears to be independent of the experimental conditions. The estimation variance is acceptable in almost all cases if one considers that a very limited data set is used. As expected, the estimation variance generally increases with noise and is larger with larger values with N. Yet a 10-fold value of N is not detrimental for the estimation performance. For some elements of $\mathbb{A}_{\bar{x},\bar{u}}$, the estimates drawn in presence of measurement noise but with all data points available (scenario B) are by far the most uncertain. This exception is rather counterintuitive and deserves more investigation. Finally, unreported results comparing constrained and unconstrained estimation show that the stability constraint becomes active in roughly 10% of the estimation runs, the latter rate being larger in the presence of noise and undersampling.

7 Conclusions and perspectives

We investigated the problem of genetic network structure identification in a stochastic hybrid modelling framework. We considered a piecewise deterministic model of genetic networks where protein synthesis is triggered by discrete random binding events and follows simple deterministic kinetics. We showed how to approximate the stochastic hybrid system locally via a linear stochastic system by considering the second order moments in stationarity. Using this approximation, we introduced an identification procedure that is based on matching the covariance function of the model to the data and provides an estimate of the average effect of each transcription factor on every gene. Extensions of the method were also discussed and include the estimation of the number of the genes in the network. Numerical results on simulated data witness the validity of the approach even in the presence of noisy and undersampled measurements. We are currently investigating on how to relax some of the assumptions mentioned in the paper and how to exploit system perturbations and experimental design to gain a more detailed insight into the structure of the network. In addition, we believe that our previous results on parameter estimation in stochastic hybrid models with known structure can be combined with the local structure identification procedure described in this paper to devise a full-blown stochastic hybrid model identification methodology. Heuristics for achieving this integration are currently under study.

Acknowledgements

This work was supported in part by the SystemsX.ch research consortium under the project YeastX.

References

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

2. Alur, R., Belta, C., Ivancic, F., Kumar, V., Mintz, M., Pappas, G., Rubin, H., Schug, J.: Hybrid modeling and simulation of biological systems. In Di Benedetto, M., Sangiovanni-Vincentelli, A., eds.: Hybrid Systems: Computation and Control. Number 2034 in LNCS, Berlin, Springer–Verlag (2001) 19–32
3. de Jong, H., Gouze, J.L., Hernandez, C., Page, M., Sari, T., Geiselmann., J.: Hybrid modeling and simulation of genetic regulatory networks: A qualitative approach. In Maler, O., Pnueli, A., eds.: Hybrid Systems: Computation and Control. Number 2623 in LNCS, Berlin, Springer–Verlag (2003) 267–282
4. Drulhe, S., Ferrari-Trecate, G., de Jong, H., Viari., A.: Reconstruction of switching thresholds in piecewise-affne models of genetic regulatory networks. In Hespanha, J., Tiwari, A., eds.: Hybrid Systems: Computation and Control. Number 3927 in LNCS, Berlin, Springer–Verlag (2006) 184–199
5. Batt, G., Ropers, D., de Jong, H., Geiselmann, J., Mateescu, R., Page, M., Schneider, D.: Validation of qualitative models of genetic regulatory networks by model checking: Analysis of the nutritional stress response in escherichia coli. Bioinformatics **21(1)** (2005) i19–i28
6. Ghosh, R., Tomlin, C.: Symbolic reachable set computation of piecewise affine hybrid automata and its application to biological modeling: Delta-notch protein signaling. IET Systems Biology **1**(1) (2004) 170–183
7. Longo, D., Hasty, J.: Dynamics of single-cell gene expression. Molecular Systems Biology **2** (2006)
8. Elowitz, M.B., Levine, A.J., Siggia, E.D., Swain, P.S.: Stochastic gene expression in a single cell. Science **297**(5584) (2002) 1183–1186
9. McAdams, H.H., Arkin, A.: It's a noisy business! genetic regulation at the nanomolar scale. Trends in Genetics **15**(2) (2002) 65–69
10. Paulsson, J.: Models of stochastic gene expression. Physics of Life Reviews **2**(2) (2005) 157–175
11. Samad, H.E., Khammash, M., Petzold, L., Gillespie, D.: Stochastic modeling of gene regulatory networks. International Journal of Robust Nonlinear Control **15** (2005) 691711
12. Cinquemani, E., Milias-Argeitis, A., Summers, S., Lygeros, J.: Stochastic dynamics of genetic networks: modelling and parameter identification. Bioinformatics **24**(23) (December 2008) 2748–2754
13. Zeiser, S., Franz, U., Wittich, O., Liebscher, V.: Simulation of genetic networks modelled by piecewise deterministic markov processes. IET Systems Biology **2** (may 2008) 113–135
14. Perkins, T., Hallett, M., Glass, L.: Inferring models of gene expression dynamics. Journal of Theoretical Biology **230**(3) (2004) 289–299
15. Fujarewicz, K., Kimmel, M., Swierniak, A.: On fitting of mathematical models of cell signaling pathways using adjoint systems. Mathematical Biosciences and Engineering **2**(3) (2005) 527–534
16. Dunlop, M., Franco, E., Murray, R.M.: A multi-model approach to identification of biosynthetic pathways. In: Proceedings of the 26th American Control Conference. (2007)
17. Cinquemani, E., Porreca, R., Ferrari-Trecate, G., Lygeros, J.: Subtilin production by bacillus subtilis: Stochastic hybrid models and parameter identification. IEEE Transactions on Automatic Control, Special Issue on Systems Biology **53** (2008) 38–50
18. Reinker, S., Altman, R., Timmer, J.: Parameter estimation in stochastic biochemical reactions. IET Systems Biology **153** (jul 2006) 168–178

19. Tian, T., Xu, S., Gao, J., Burrage, K.: Simulated maximum likelihood method for estimating kinetic rates in gene expression. Bioinformatics **23**(1) (2007) 84–91
20. Golightly, A., Wilkinson, D.: Bayesian inference for stochastic kinetic models using a diffusion approximation. Biometrics (61) (sept 2005) 781788
21. Zavlanos, M.M., Julius, A., Boyd, S.P., Pappas, G.J.: Identification of stable genetic networks using convex programming. In: Proceedings of the American Control Conference, Seattle, WA (June 2008)
22. Bansal, M., Belcastro, V., Ambesi-Impiombato, A., di Bernardo, D.: How to infer gene networks from expression profiles. Molecular Systems Biology **3**(78)
23. Gardner, T.S., di Bernardo, D., Lorenz, D., Collins, J.J.: Inferring Genetic Networks and Identifying Compound Mode of Action via Expression Profiling. Science **301**(5629) (2003) 102–105
24. van Overschee, P., De Moor, B.L.: Subspace Identification for Linear Systems: Theory - Implementation - Applications. Springer (1996)
25. Golding, I., Paulsson, J., Zawilski, S.M., Cox, E.C.: Real-time kinetics of gene activity in individual bacteria. Cell **123**(6) (2005) 1025–1036
26. Cai, L., Friedman, N., Xie, X.S.: Stochastic protein expression in individual cells at the single molecule level. Nature **440** (mar 2006) 358–362
27. Davis, M.: Piecewise-deterministic Markov processes: A general class of non-diffusion stochastic models. Journal of the Royal Statistical Society **B 46**(3) (1984) 353–388
28. Boyd, S.P., Vandenberghe, L.: Convex optimization. Cambridge University Press (2004)
29. Lacy, S.L., Bernstein, D.S.: Subspace identification with guaranteed stability using constrained optimization. IEEE Transactions on Automatic Control **48**(7) (2003)
30. I.Smith, M.: A Schur algorithm for computing matrix pth roots. SIAM Journal on Matrix Analysis and Applications **24**(4) (2003) 971–989
31. Bini, D.A., Higham, N.J., Meini, B.: Algorithms for the matrix pth root. Numerical Algorithms **39** (2005) 349–378

A Proofs

Proof of Proposition 2. The equation for $\bar{x} = \mathbb{E}[x(t)] = \mathbb{E}[x(t+T)]$ follows from $\mathbb{E}[x(t+T)] = \mathbb{E}\big[Ax(t) + g(\bar{u}) + G_{\bar{u}}\big(u(t+T) - \bar{u}\big)\big] = A\mathbb{E}[x(t)] + g(\bar{u})$. Using this equation and Eq. (4), $\tilde{x}(t+T) = Ax(t) + g(\bar{u}) + G_{\bar{u}}\tilde{u}(t+T) - \bar{x} = Ax(t) + G_{\bar{u}}\tilde{u}(t) - A\bar{x}(t)$, which is (5).

Proof of Proposition 4. *Without loss of generality, we shall prove the result for* $T = 1$. From Assumption 3, $\mathbb{E}[\tilde{u}(t+1)|x(t), x(t-\ell)] = \mathbb{E}[\tilde{u}(t+1)|x(t)]$ for all $\ell > 0$, where $\mathbb{E}[\cdot|\cdot]$ denotes conditional expectation. Eq. (7) is given by $\mathbb{E}[\tilde{x}(t+\ell+1)\tilde{x}(t)^T] = A\mathbb{E}[\tilde{x}(t+\ell)\tilde{x}(t)^T] + G_{\bar{u}}\mathbb{E}[\tilde{u}(t+\ell+1)\tilde{x}(t)^T]$ where

$$\begin{aligned}\mathbb{E}[\tilde{u}(t+\ell+1)\tilde{x}(t)^T] &= \mathbb{E}\big[\mathbb{E}[\tilde{u}(t+\ell+1)\tilde{x}(t)^T|x(t+\ell), x(t)]\big] = \\ &= \mathbb{E}\big[\mathbb{E}[\tilde{u}(t+\ell+1)|x(t+\ell)]\tilde{x}(t)^T\big] = F_{\bar{x}}\mathbb{E}[\tilde{x}(t+\ell)\tilde{x}(t)^T\big].\end{aligned}$$

To get Eq. (8), note that $\Sigma_x(0) = \mathbb{E}[\tilde{x}(t)\tilde{x}(t)^T] = \mathbb{E}[\tilde{x}(t+1)\tilde{x}(t+1)^T]$. Using (6) to expand the product in the latter expectation yields

$$\Sigma_x(0) = A\Sigma_x(0)A^T + G_{\bar{u}}\mathbb{E}[\tilde{u}(t+1)\tilde{x}(t)^T]A^T + A\mathbb{E}[\tilde{x}(t)\tilde{u}(t+1)^T]G_{\bar{u}}^T + \\ + G_{\bar{u}}\mathbb{E}[\tilde{u}(t+1)\tilde{u}(t+1)^T]G_{\bar{u}}^T. \quad (12)$$

In turn, $\mathbb{E}[\tilde{u}(t+1)\tilde{x}(t)^T] = \mathbb{E}\big[\mathbb{E}[\tilde{u}(t+1)\tilde{x}(t)^T|x(t)]\big] = F_{\bar{x}}\mathbb{E}[\tilde{x}(t)\tilde{x}(t)^T] = F_{\bar{x}}\Sigma_x(0)$ and (writing $\tilde{u}$ for $\tilde{u}(t+T)$ and $\tilde{x}$ for $\tilde{x}(t)$)

$$\begin{aligned}
\mathbb{E}[\tilde{u}\tilde{u}^T] &= \mathbb{E}\Big[\big(u - \mathbb{E}[u|x] + \mathbb{E}[u|x] - \bar{u}\big)\big(u - \mathbb{E}[u|x] + \mathbb{E}[u|x] - \bar{u}\big)^T\Big] \\
&= \mathbb{E}\Big\{\mathbb{E}\Big[\big(u - \mathbb{E}[u|x] + \mathbb{E}[u|x] - \bar{u}\big)\big(u - \mathbb{E}[u|x] + \mathbb{E}[u|x] - \bar{u}\big)^T|x\Big]\Big\} \\
&\overset{*}{=} \mathbb{E}\Big\{\mathbb{E}\Big[\big(u - \mathbb{E}[u|x]\big)\big(u - \mathbb{E}[u|x]\big)^T|x\Big] + \mathbb{E}\Big[\big(\mathbb{E}[u|x] - \bar{u}\big)\big(\mathbb{E}[u|x] - \bar{u}\big)^T|x\Big]\Big\} \\
&= \mathbb{E}\big[\mathrm{Var}(u|x)\big] + \mathbb{E}\Big[\mathbb{E}\big[(F_{\bar{x}}\tilde{x})(F_{\bar{x}}\tilde{x})^T|x\big]\Big] = Q + F_{\bar{x}}\Sigma_x(0)F_{\bar{x}}^T.
\end{aligned}$$

(To verify "$*$", expand the product in the LHS and note that, since $\mathbb{E}[u|x] - \bar{u}$ is constant for given x and $\mathbb{E}\big[u - \mathbb{E}[u|x]|x\big] = 0$, the cross-product $\mathbb{E}\big[\big(u - \mathbb{E}[u|x]\big)\big(\mathbb{E}[u|x] - \bar{u}\big)^T|x\big]$ vanishes and so does its transpose.) Substituting these equations into (12) and rearranging the terms yields the result.

Proof of Proposition 5. The result can be deduced from the representation (9). *Without loss of generality, we may restrict to the case* $N = 1$. For the sake of simplicity let us also drop linearization points $\bar{x}$ and $\bar{u}$ from the notation. For $t \in \mathscr{T}$ and every $\ell \geq 1$,

$$\begin{aligned}
\mathbb{E}\big[\tilde{x}(t + \ell T + T)\tilde{y}(t)^T\big] &= \mathbb{E}\big[\big(\mathbb{A}\tilde{x}(t + \ell T) + Gw(t + \ell T)\big)\tilde{y}(t)^T\big] \\
&= \mathbb{A}\mathbb{E}\big[\tilde{x}(t + \ell T)\tilde{y}(t)^T\big].
\end{aligned}$$

On the other hand, $\mathbb{E}\big[\tilde{y}(t + \ell T)\tilde{y}(t)^T\big] = \mathbb{E}\big[\tilde{x}(t + \ell T)\tilde{y}(t)^T\big]$ because $n(t + \ell T)$ is uncorrelated with $\tilde{y}(t)$. Therefore $\Lambda(\ell + 1) = \mathbb{A}\Lambda(\ell)$. For $\ell = 0$,

$$\mathbb{E}\big[\tilde{x}(t + T)\tilde{y}(t)^T\big] = \mathbb{E}\big[\big(\mathbb{A}\tilde{x}(t) + Gw(t)\big)\tilde{y}(t)^T\big] = \mathbb{A}\mathbb{E}\big[\tilde{x}(t)\tilde{y}(t)^T\big],$$

where $\mathbb{E}\big[\tilde{x}(t)\tilde{y}(t)^T\big] = \mathbb{E}\big[\big(\tilde{y}(t) - n(t)\big)\tilde{y}(t)^T\big] = \mathbb{E}\big[\tilde{y}(t)\tilde{y}(t)^T\big] - \mathbb{E}\big[n(t)n(t)^T\big]$, hence $\Lambda(1) = \mathbb{A}(\Lambda(0) - R)$.