Inference of the statistics of a modulated promoter process from population snapshot gene expression data [⋆]

Eugenio Cinquemani [*]

[*] *Univ. Grenoble Alpes, Inria, 38000 Grenoble, France*
(e-mail: eugenio.cinquemani@inria.fr).

Abstract: In previous work, we have developed mathematical tools for the analysis of single-cell gene expression data from population snapshots, and an inference algorithm for the estimation of stationary statistics of promoter activation. In this work, we address the inference problem in the nonstationary case of modulated processes. This is of special relevance to control scenarios, where an exogenous input modulates the time evolution of promoter activation. We provide an effective method for the computation of the output statistics of a reaction network with a nonstationary, causal input process of modulated form. Based on this we devise and demonstrate an algorithm for the reconstruction of the promoter (input) process statistics from snapshot data.

Keywords: Reporter gene systems, Controlled Markov chains, Regularized estimation, Splines

1. INTRODUCTION

Noise in gene expression is at the roots of important phenomena in cell life and evolution, such as bet-hedging and the emergence of specific gene regulatory patterns (Raj and van Oudenaarden, 2008; Rao et al., 2002). Its characterization is thus key for understanding cellular dynamics, and it has been the object of intense investigation (Thattai and van Oudenaarden, 2001; Paulsson, 2005; Swain et al., 2002; Kaern et al., 2005) . Among the frontiers of systems and synthetic biology is control of gene expression (Uhlendorf et al., 2012; Chait et al., 2017; Milias-Argeitis et al., 2016). A primary objective of control is the exploration of cellular dynamics. In particular, control enables a deeper investigation of gene expression noise. This calls for the development of statistical analysis and inference methods (Komorowski et al., 2009; Munsky et al., 2009; Neuert et al., 2013; Ocone et al., 2015) where the presence of control is explicitly accounted for.

In previous work (Cinquemani, 2019), we have addressed inference of promoter activity statistics from gene expression population-snapshot data, that is, measurements of the statistical distribution of the gene expression product in samples from the cellular population collected at different times (Hasenauer et al., 2011). We framed the problem in the context of (first-order) stochastic reaction networks with a causal input on reaction rates, and developed Generalized Moment Equations (GMEs), a marginalized form of the well-known Moment Equations (MEs, see *e.g.* Lestas et al. (2008)) that describes the statistics of the network state process as a function of the input process statistics. We then used the GMEs to devise a method to reconstruct, in particular, the autocovariance function of an arbitrary stationary promoter process. Unfortunately, the method does not apply to controlled gene expression processes, due to the typical nonstationarity of the control action.

In this work, we propose a generalization of the methods in Cinquemani (2019) to modulated input processes (Zhao and Li, 2013). With this terminology we refer to stationary processes reshaped by a time-varying signal. In their matrix formulation, modulated processes account for a variety of scenarios of special relevance to gene expression, where a control input acts as an additive or multipicative factor on random gene expression noise (see *e.g.* Munsky et al. (2009); Uhlendorf et al. (2012); Milias-Argeitis et al. (2016)). Based on a very flexible characterization of modulating functions via splines, we develop a strategy for the efficient numerical solution of GMEs in presence of modulated inputs. Then, we extend the inference method in Cinquemani (2019) to the estimation of modulated process statistics, and discuss reconstruction performance and the role of the modulating input via numerical simulations.

GMEs are reviewed in Sec. 2. The inference problem and method developed for stationary promoter processes in discussed is Sec. 3. Modulated input processes as well as an effective solution of the corresponding GMEs are presented in Sec. 4. The novel, generalized inference method is then presented in the same section. Numerical simulations demonstrating the method are in Sec. 5. Finally, conclulsions are drawn in Sec. 6.

Notation: $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{R}$ and $\mathbb{R}_+$ denote natural, integer, real and nonnegative real numbers, respectively. For a set $T \subset \mathbb{R}$, $\mathbb{1}_T(\cdot)$ is the indicator function of T, and $\mathbb{1}(\cdot)$ is the unit step function $\mathbb{1}_{[0,+\infty)}(\cdot)$. $||\cdot||$ denotes Euclidean norm. For two random vectors X and Y, $\mathbb{E}[X]$ denotes expectation of X, $\mathrm{Cov}(X,Y) = \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])^T]$ (superscript "^{T}" denoting transposition) and $\mathrm{Var}(X) = \mathrm{Cov}(X,X)$. $\mathbb{P}[\,\cdot\,]$ denotes probability.

[⋆] This work was funded in part by the French national research agency (ANR) via project MEMIP (ANR-16-CE33-0018), and by the Inria IPL CoSy. The author thanks Jakob Ruess for useful discussions.

2. GENERALIZED MOMENT EQUATIONS

A reaction network is a family of $n \in \mathbb{N}$ chemical species and $m \in \mathbb{N}$ reactions that may occur among them in a given reaction volume. At a time t, let $X_i(t)$, with $i = 1, \ldots, n$, be the number of molecules of the ith species. Let S_j, with $j = 1, \ldots, m$, be the column vector whose ith row is the net change in X_i when reaction j occurs. $S = [S_1 \ \cdots \ S_m] \in \mathbb{Z}^{n\times m}$ is the network stoichiometry matrix. Assuming a well-stirred reaction volume, $X = [X_1 \ \cdots \ X_n]^T$ can be modelled as a Markov process, with

$$\mathbb{P}[X(t+dt) = x + S_j | X(t) = x] = w_j(x,t)dt + o(dt),$$

where the functional form of the reaction rates (propensities) $w(x,t) = [w_1(x,t) \ \cdots \ w_m(x,t)]^T$ is typically fixed by the mass-action laws (Gillespie, 1992), and $o(dt)$ tends to 0 faster than dt. We consider the case where rates are affine functions of the state,

$$w(X,t) = WX(t) + F(t), \qquad (1)$$

with $W \in \mathbb{R}_+^{m\times n}$ a constant matrix and $F(t)$ some exogenous input taking values in $\mathbb{R}_+^m$. This captures zeroth- and first-order reactions and suffices to describe many biochemical processes of interest (notably gene expression) in an exact or approximate manner (Kaern et al., 2005; Thattai and van Oudenaarden, 2001). Depending on the context, F may represent environmental perturbations, intracellular regulation, or control.

Let $\mu(t) = \mathbb{E}[X(t)]$, $\Sigma(t) = \mathrm{Var}\big(X(t)\big)$. Given S, functions $w(\cdot,\cdot)$ and a deterministic function $F(\cdot)$, the time dynamics of μ and Σ (and of higher-order moments) are described by a well-known set of Ordinary Differential Equations (ODEs) called the Moment Equations (MEs), see *e.g.* Lestas et al. (2008). For the more general case considered in this paper, where $F(t)$ is a stochastic (vector) process with well-defined second-order moments, let us further define $\mu_F(t) = \mathbb{E}[F(t)]$, $\rho_F(t,\tau) = \mathrm{Cov}\big(F(t), F(\tau)\big)$, $\xi_F(t) = \mathrm{Cov}\big(X(0), F(t)\big)$. In Cinquemani (2019), we proved in particular the following result.

Proposition 1. Assume that first- and second-order moments of F are uniformly bounded. Then, for $t \geq 0$,

$$\dot{\mu}(t) = SW\mu(t) + S\mu_F(t), \qquad (2)$$

$$\dot{\Sigma}(t) = SW\Sigma(t) + \Sigma(t)W^TS^T + Q(t) + V_\xi(t) + V_\xi^T(t) + V_\rho(t) + V_\rho^T(t), \qquad (3)$$

where, denoting $\ell(t) = \exp(SWt)\mathbb{1}(t)$,

$$Q(t) = S\mathrm{diag}\big(W\mu(t) + \mu_F(t)\big)S^T, \qquad (4)$$

$$V_\xi(t) = S\xi_F(t)^T\ell(t)^T, \qquad (5)$$

$$V_\rho(t) = \int_0^{+\infty} d\tau S\rho_F(t,\tau)S^T\ell(t-\tau)^T. \qquad (6)$$

We call Eq. (2)–(3) the Generalized Moment Equations (GMEs). For F a deterministic profile (that is, $F = \mu_F$, $\rho_F = 0$, $\xi_F = 0$), they reduce to the standard MEs. Note that the result holds irrespective of F being Markovian or stationary, however, a form of stochastic causality between the input process F and the state process X is subsumed by the definition of the rate functions (1) (see Cinquemani (2019)). Crucially, GMEs are linear in the input statistics μ_F, ρ_F and ξ_F. As discussed in the sequel, this fact enables the development of efficient methods for practical applications.

3. INFERENCE OF PROMOTER STATISTICS

At the single-cell level, gene expression is well described by the so-called random telegraph model, *i.e.* the reaction network defined by the four stochastic reactions

$$\varnothing \xrightarrow{k_M\cdot U} M, \qquad M \xrightarrow{d_M} \varnothing, \qquad M \xrightarrow{k_P} M + P, \quad P \xrightarrow{d_P} \varnothing \qquad (7)$$

(Kaern et al., 2005; Paulsson, 2005), where M and P denote mRNA and protein species, respectively, while $\theta = (k_M, d_M, k_P, d_P)$ are positive rate parameters. The binary random process U describes the state of the promoter that controls gene expression. Let X_1 and X_2 be the number of mRNA and protein molecules, respectively. Ordering reactions (7) from left to right, then top to bottom,

$$S = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}.$$

Moreover, the rate equations are as in (1), with

$$W = \begin{bmatrix} 0 & 0 \\ d_M & 0 \\ k_P & 0 \\ 0 & d_P \end{bmatrix}, \qquad F(t) = KU(t), \qquad K = \begin{bmatrix} k_M \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

The random outcomes of the joint process (U, X) may be thought of as different gene expression time profiles in different cells of a genetically homogeneous population. Randomness of U accounts in particular for different activation profiles and can be interpreted as extrinsic noise (Swain et al., 2002).

In order to study the expression of a gene of interest, it is possible to engineer cells such that P is a fluorescent protein (see *e.g.* de Jong et al. (2010)). Then, individual-cell fluorescent protein levels $X_2(t)$ can be quantified by means of videomicroscopy or flow-cytometry. In particular, this allows one to get population snapshots of gene expression distribution at different times. We focus on the empirical mean and variance of gene expression snapshots, which provide noisy measurements of the mean and variance of X_2. Denoting these measurements with $\tilde{\mu}_k$ and $\tilde{\sigma}_k^2$, for a sequence of measurement times t_k with $k = 1, \ldots, M$,

$$\tilde{\mu}_k = C\mu(t_k) + e_k^\mu, \qquad (8)$$

$$\tilde{\sigma}_k^2 = C^T\Sigma(t_k)C + e_k^\sigma, \qquad (9)$$

where $C = [0 \ 1]^T$ selects from μ and Σ the mean and variance of the protein count X_2 (the proportionality factor relating X_2 with observed fluorescence is ignored for simplicity). Measurement errors e_k^μ and e_k^σ can be statistically characterized (Zechner et al., 2012).

3.1 Inference of the statistics of a stationary process

On the basis of the gene expression model above (with known parameters θ), in Cinquemani (2019), a method for the reconstruction of the second-order statistics of the promoter state process U from snapshot data was developed for the case where U is stationary. In particular, under a few assumptions (notably $X(0) = 0$, as met by appropriate experiments, which implies $V_\xi = 0$ in (3)) the method leverages the linearity of the GMEs (2)–(3) in μ_F and ρ_F to infer the (scalar-valued) stationary autocovariance function $\bar{\rho}_U(\cdot) = \mathrm{Cov}\big(U(t+\cdot)U(t)\big)$ from measurements (9), without any mechanistic model for U. The interest of this problem is that the autocovariance function carries information about hidden regulatory processes, and its reconstruction from data is thus an important first step in characterization of gene expression. Notice that

$$\mu_F(t) = K\bar{\mu}_U, \qquad \rho_F(t+\delta, t) = K\bar{\rho}_U(\delta)K^T, \tag{10}$$

so that linearity of (2)–(3) in μ_F and ρ_F implies their linearity in $\bar{\mu}_U$ and $\bar{\rho}_U$. Assuming identification of the unknown constant $\bar{\mu}_U$ from the mean data (8) and Eq. (2) in a preliminary step, the method seeks the autocovariance function $\bar{\rho}_U$ such that the corresponding solution of (3) best fits the measurements (9) in a Least-Squares (LS) sense. This is done via a finite expansion of the unknown autocovariance function. Let $\mathscr{R}(\delta) = [r_1(\delta), \ldots, r_N(\delta)]^T$ be N appropriately chosen functions such that the convex cone generated by $\mathscr{R}$ approximates well the space of stationary autocovariance functions. Then, for some vector of nonnegative coefficients $c = [c_1, \ldots, c_N]^T$, the equation

$$\bar{\rho}_U(\delta) = c_1 r_1(\delta) + \ldots + c_N r_N(\delta) = \mathscr{R}(\delta) \cdot c \tag{11}$$

holds with good approximation at all time lags δ. For inference purposes, $\mathscr{R}$ is fixed *a priori*, and c is the quantity to be determined from the data. For any $t \geq 0$, let $\hat{\Sigma}(t|\bar{\rho}_U, \bar{\mu}_U)$ denote the solution $\Sigma(t)$ of the GMEs (2)–(3), with μ_F and ρ_F as in (10), as a function of $\bar{\mu}_U$ and $\bar{\rho}_U$. Define $\mathscr{V}(\cdot) = [v_1(\cdot), \ldots, v_N(\cdot)]$ where, for every l, $v_l(\cdot) = C^T\hat{\Sigma}(\cdot\,|r_l, 0)C$, and $v_0(\cdot) = C^T\hat{\Sigma}(\cdot\,|0, \bar{\mu}_U)C$. Using (11), in view of the linearity of the GMEs, it holds that $C^T\hat{\Sigma}(t|\bar{\rho}_U, \bar{\mu}_U)C = v_0 + \mathscr{V}(t) \cdot c$. Therefore, the LS estimate $\hat{c}$ of c from data (9) can be written as

$$\inf_{c \in \mathbb{R}^N_+} \sum_{k=1}^{M} \alpha_k^2 \left(\tilde{\sigma}^2_Y(t_k) - v_0(t_k) - \mathscr{V}(t_k)c\right)^2, \tag{12}$$

where the weights $\alpha_k = \mathrm{std}(e^\sigma_k)^{-1}$ are known (Zechner et al., 2012). The estimate of $\bar{\rho}_U$ is then defined as $\hat{\rho}_U = \mathscr{R} \cdot \hat{c}$. In Cinquemani (2019), additional constraints and a regularization term are introduced to ensure that the problem is well-posed (De Nicolao et al., 1997) and that $\hat{\rho}_U$ is a well-defined autocovariance function, while preserving convexity of the optimization problem.

The viability of the method crucially depends on the ability to compute the transforms $\mathscr{V}$ of the approximating functions $\mathscr{R}$. For a convenient choice of univariate, scalar-valued functions r_l (notably, indicator functions), the integral term $V_{r_l}(t)$ (see Eq. (6)) can be written out as an explicit function of t. In turn, for every l, this allows one to calculate the basis function transform $v_l(t_k)$ at all times t_k by standard numerical integration of (3) (with r_l in place of $\bar{\rho}_U$), leading to an efficient numerical implementation of the inference method (see Cinquemani (2019) for details).

4. EXTENSION TO MODULATED PROCESSES

We now want to pursue the efficient solution of the GMEs and a method for reconstruction of input process statistics from snapshot data in the case where the input process F is subjected to a control signal. We look at a specific class of nonstationary processes, which we refer to as modulated processes (Zhao and Li, 2013):

$$F(t) = G(t)E(t), \tag{13}$$

where $E(t)$ is a second-order stationary (vector) process taking values in $\mathbb{R}^q$ and $G(t)$ is a deterministic (matrix) function of time taking values in $\mathbb{R}^{m\times q}$. If $\bar{\mu}_E$ and $\bar{\rho}_E(\cdot)$ are respectively the stationary mean and stationary autocovariance function of E, then

$$\mu_F(t) = G(t)\bar{\mu}_E, \quad \rho_F(t,\tau) = G(t)\bar{\rho}_E(t-\tau)G(\tau)^T. \tag{14}$$

This case is of relevance, in particular, for gene expression models. Indeed, thanks to its general matrix-vector form, Eq. (13) includes as special cases additional and multiplicative forms of promoter regulation and control, acting at the level of basal expression rate or in terms of expression strength (see *e.g.* Chait et al. (2017); Berthoumieux et al. (2013); Fiore et al. (2016)).

Here, we limit ourselves to the case where the control signal $G(t)$ is known. For simplicity we focus on the case where E is a scalar process ($q = 1$). The approach we take is analogous to the one reviewed in the previous section. For a given $G(t)$, we aim at reconstructing the stationary autocovariance $\bar{\rho}_E$ from variance measurements (9) by a LS approach, leveraging the GMEs (2)–(3) and (14). (As above, we assume that the constant $\bar{\mu}_E$ is identified from mean data in a preliminary step). From an estimate of $\bar{\rho}_E$, an estimate of the nonstationary ρ_F follows via (14).

Given functions $\mathscr{R}$ for the finite expansion of $\bar{\rho}_E$, the main challenge in the new scenario is the computation of the transformed functions $\mathscr{V}$. Formally, these are still given by $v_l = C^T\hat{\Sigma}(\cdot\,|r_l, 0)C$, where, in the light of (14), $\hat{\Sigma}(\cdot\,|r, m)$ now expresses the solution of (2)–(3) relative to a hypothetical mean m and autocovariance function r of E. In practice, though, calculating $\hat{\Sigma}(\cdot\,|r, m)$ by numerical integration of (2)–(3) requires evauation at all times t of the integral term (6), which is now

$$V_\rho(t) = \int_0^{+\infty} d\tau SG(t)\bar{\rho}_E(t-\tau)G(\tau)^T S^T \ell(t-\tau)^T. \tag{15}$$

For a generic $G(t)$, this term cannot be written as an explicit function of t. To solve the GMEs, the idea is then to express V_ρ itself as the solution of a system of differential equations to be numerically solved jointly with (2)–(3). For convenient classes of functions $G(t)$ (possibly used as approximations of more general functions) this is the subject of the section that follows.

4.1 Computation of the approximating function transforms

For a modulated process (13), eliminating τ from (15) by the change of variable $\delta = t - \tau$, one finds that, for a generic function r in place of $\bar{\rho}_E$,

$$\begin{aligned} V_r(t) &= SG(t)\int_0^t d\delta r(\delta)G(t-\delta)^T S^T \ell(\delta)^T \\ &= SG(t)H_{0,r}(t), \end{aligned} \tag{16}$$

where $H_{0,r}(t)$ is defined as the integral in the first line. We aim at giving a differential expression to $H_{0,r}(t)$. For later use, for indices i such that the ith derivative $G^{(i)}(\cdot)$ of $G(\cdot)$ exists, define

$$H_{i,r}(t) = \int_0^t d\delta r(\delta)G^{(i)}(t-\delta)^T S^T \ell(\delta)^T, \tag{17}$$

which is also consistent with the above definition of $H_{0,r}$. For ease of notation, we will now temporarily write H_i in place of $H_{i,r}$.

Proposition 2. Suppose that $G(t)$ is a matrix of polynomials of order up to $d-1$. Then, for $t \geq 0$,

$$\dot{H}_i(t) = r(t)G^{(i)}(0)^T S^T \ell(t)^T + H_{i+1}(t), \quad H_i(0) = 0, \tag{18}$$

with $i = 0, 1, \ldots, d-2$, and

$$\dot{H}_{d-1}(t) = r(t)G^{(d-1)}(0)^T S^T \ell(t)^T, \quad H_{d-1}(0) = 0. \tag{19}$$

Proof: For any i, the derivative of (17) is $[r(\delta)G^{(i)}(t-\delta)^T S^T \ell(\delta)^T]_{\delta=t} + \int_0^t d\delta \frac{d}{dt}[r(\delta)G^{(i)}(t-\delta)^T S^T \ell(\delta)^T]$, which is (18). For $i = d-1$, since $G^{(d)}(t) = 0$ for all t, the above gives (19). The initial conditions $H_i(0) = 0$ are an immediate consequence of (17).

Eq. (18)–(19) constitute an ODE system for $H_0(t)$, as desired. In practice, though, the utility of this result is limited. In general $G(t)$ may not be made of polynomials, and a polynomial approximation of a function over an interval generally requires polynomials of high degree. This would lead to a blow-up of the differential system size, which increases with d. It turns out that the above result can be extended to splines, a class of piecewise polynomial functions that are well suited to function approximation. In particular, suppose that the entries of G are piecewise polynomials of order up to $d-1$, with continuous derivatives up to $G^{(d-2)}$ (Wahba, 1990). Let the knots (points of junction of the different polynomials) be placed at $T_1, \ldots, T_p$, with $T_1 < T_2 < \ldots < T_p$ (without loss of generality, we assume all entries of G to have equally placed knots). Notice that, for some constant matrices G_j,

$$G^{(d-1)}(\tau) = \begin{cases} G_0, & \tau < T_1, \\ G_j, & \tau \in [T_j, T_{j+1}),\ j = 1, \ldots, p-1, \\ G_p, & \tau \geq T_p. \end{cases} \tag{20}$$

Proposition 3. Suppose that the entries of $G(t)$ are piecewise polynomials as defined above. Then, for $t \geq 0$ and $i = 0, 1, \ldots, d-2$, Eq. (18) holds. Moreover,

$$\dot{H}_{d-1}(t) = H_0^+(t) + \sum_{j=1}^{p-1} \left(H_j^+(t) - H_{j-1}^-(t) \right), \tag{21}$$

with $H_{d-1}(0) = 0$, where, for all relevant j,

$$\begin{aligned} H_j^+(t) &= r(t-T_j)G_j^T S^T \ell(t-T_j)^T, \\ H_j^-(t) &= r(t-T_{j+1})G_j^T S^T \ell(t-T_{j+1})^T. \end{aligned} \tag{22}$$

Proof: The validity of Eq. (18) for $i = 0, 1, \ldots, d-2$ is proven as in Prop. 2, also in view of the continuity of the derivatives of G up to $G^{(d-2)}$. Next, using Eq. (20), simple manipulations of Eq. (17) allow one to expand the expression of $H_{d-1}(t)$ as

$$\int_{t-T_1}^{t} d\delta r(\delta)G_0^T S^T \ell(\delta)^T + \sum_{j=1}^{p-1} \int_{t-T_{j+1}}^{t-T_j} d\delta r(\delta) G_j^T S^T \ell(\delta)^T + \int_0^{t-T_p} d\delta r(\delta) G_p^T S^T \ell(\delta)^T$$

(since $\ell(\delta) = 0$ for $\delta < 0$, this holds for any value of t). Taking the derivative in t term by term yields

$$H_0^+(t) - H_0^-(t) + \sum_{j=1}^{p-1} \left(H_j^+(t) - H_j^-(t) \right) + H_p^+(t).$$

Eq. (21) is a straightforward rearrangement of the above.

Notice that the generic jth term of the summation in (21) is nonzero only for $t > T_j$. The generalization of (21) to an unbounded number of knots over an infinite time horizon is thus straightforward. In conclusion, for an arbitrary (measurable) function r, Eq. (18) (with $i = 0, \ldots, d-2$) and (21) jointly give a system of ODEs for the computation of $H_{0,r}(\cdot)$. For any time t, $V_r(t)$ then follows from (16).

4.2 Inference method

We are now ready to state our method for the inference of $\bar{\rho}_E$ and ρ_F from measurements (9), for a known function G. For the sake of this paper, we assume that E is a scalar (stationary) process, and $\mathscr{R} = [r_1 \cdots r_N]$ is a collection of indicator functions, though generalizations are immediate. As in Cinquemani (2019), the method assumes $X(0) = 0$. Parameters S and W are assumed known and, as explained above, $\bar{\mu}_E$ is identified in advance.

(i) (Computation of the v_l) Perform numerical integration of

$$\begin{aligned} \dot{\mu}(t) &= SW\mu(t) + SG(t)\bar{\mu}_E, \\ \dot{\Sigma}(t) &= SW\Sigma(t) + \Sigma(t)W^T S^T + \\ &\qquad S\mathrm{diag}\big(W\mu(t) + G(t)\bar{\mu}_E\big)S^T, \end{aligned}$$

with $\mu(0) = 0$ and $\Sigma(0) = 0$ to get the solution $\hat{\Sigma}(t_k|0, \bar{\mu}_E)$ at all measurement times, and set $v_0(t_k) = C^T \hat{\Sigma}(t_k|0, \bar{\mu}_E)C$, with $k = 1, \ldots, M$. Next, for every function r_l, with $l = 1, \ldots, N$, perform numerical integration of the augmented ODE system

$$\begin{aligned} \dot{\Sigma}(t) &= SW\Sigma(t) + \Sigma(t)W^T S^T + \\ &\qquad SG(t)H_{0,r_l}(t) + H_{0,r_l}(t)^T G(t)^T S^T, \\ \dot{H}_{0,r_l}(t) &= r_l(t)G^{(0)}(0)^T S^T \ell(t)^T + H_{1,r_l}(t), \\ &\ \ \vdots \\ \dot{H}_{d-2,r_l}(t) &= r_l(t)G^{(d-2)}(0)^T S^T \ell(t)^T + H_{d-1,r_l}(t), \\ \dot{H}_{d-1,r_l}(t) &= H_{0,r_l}^+(t) + \sum_{j=1}^{p-1} \left(H_{j,r_l}^+(t) - H_{j-1,r_l}^-(t) \right), \end{aligned}$$

with $H_{0,r_l}(0) = \ldots = H_{d-1,r_l}(0) = 0$ and $\Sigma(0) = 0$, to get the solution $\hat{\Sigma}(t_k|r_l, 0)$ at all measurement times. Set $\mathscr{V}(t_k) = [v_1(t_k), \ldots, v_N(t_k)]$, where $v_l(t_k) = C^T \hat{\Sigma}(t_k|r_l, 0)C$, with $l = 1, \ldots, N$ and $k = 1, \ldots, M$.

(ii) (Estimation of c) Compute $\hat{c}$ as a solution to the convex optimization problem

$$\min_{c \in \mathbb{R}^N} \sum_{k=1}^{M} \alpha_k^2 \left(\tilde{\sigma}_Y^2(t_k) - v_0(t_k) - \mathscr{V}(t_k)c \right)^2 + \lambda \cdot c^T Q c$$

s.t. $\mathscr{T}(c) \in \mathscr{C}_N$

where $\mathscr{T}(c)$ denotes the symmetric Toeplitz matrix with first column equal to c, $\mathscr{C}_N$ is the convex cone of positive semi-definite matrices of order N, and $Q \in \mathscr{C}_N$.

(iii) (Calculation of $\hat{\rho}_E$ and $\hat{\rho}_F$) Return the estimate of $\bar{\rho}_E$ as $\hat{\rho}_E(\cdot) = \mathscr{R}(\cdot)\hat{c}$ and the estimate of ρ_F as $\hat{\rho}_F(z,t) = G(z)\hat{\rho}_E(z-t)G(t)^T$.

In (ii), the penalty term $c^T Q c$ is used to enforce regularity of the solution. For a suitable choice of Q, $c^T Q c$ accounts in particular for the average curvature of the candidate solution $\mathscr{R}c$. The optimization constraint ensures that $\mathscr{R}c$ is a well-defined (positive semi-definite) autocovariance function. Factor $\lambda \geq 0$ is a regularization weight that can be chosen automatically via repeated executions of step (ii) with different candidate values of λ. See more details in Cinquemani (2019).

5. NUMERICAL DEMONSTRATION

We now demonstrate by simulation the inference procedure of Sec. 4. We consider the gene expression model of Sec. 3, modified for the presence of a modulating signal. That is, in place of $F(t) = KU(t)$, we assume that $F(t) = G(t)E(t)$, where $G(t) = [k_M g(t)\ 0\ 0\ 0]^T$ and $E(t) = U(t)$. The parameter values $\theta = (k_M, d_M, k_P, d_P) = (0.5, 0.1, 0.2, 0.01)$, as well as the definition of the mean $\bar{\mu}_E = \bar{\mu}_U = 0.5$ and autocovariance function $\bar{\rho}_E = \bar{\rho}_U$ (shown in Fig. 1) of U are as in Cinquemani (2019). To explore the role of the (known) control input g in the estimation problem, we consider two different modulating signals, G' and G''. For G' we take $g(t) = (1 + \gamma \cdot t)^{-1}$, while for G'' we take $g(t) = 1 - (1 + \gamma \cdot t)^{-1}$ $(\gamma = 0.02)$. These control inputs are qualitatively different: G' is a monotonically decreasing signal with unit value at zero, while G'' is monotonically increasing and null at the origin (from now on, with some abuse of terminology, we refer to G' and G'' instead of their first entry as the control inputs). Toward the application of the methods of Sec. 4, we approximate G' and G'' by cubic splines $(d = 3)$, with knots at $T_j = (j - 1) \cdot T$, with $T = 1$, $j = 1, \ldots, p$ and $p = 100$. Plots of (the spline approximations of) G' and G'') are visualized in the insets of Fig. 1.

For this stochastic gene expression system, we simulate data (8) and (9) analogous to Cinquemani (2019). That is, we consider empirical mean and variance measurements collected from samples of 10^6 cells at times $t_k = 5 \cdot k$, with $k = 1, \ldots M$ and $M = 20$, with initial state $X(0) = 0$. For inference, we assume that the gene expression parameters θ are known, and so is the control input. Additionally, given the simplicity of estimating $\bar{\mu}_E$ from measurements (8), we assume the latter to be known, and focus on testing the reconstruction of autocovariance functions from measurements (9). For both G' and G'', we simulate 100 datasets, and draw estimates $\hat{\rho}_E$ and $\hat{\rho}_F$ for every dataset with the method of Sec. 4.2. Approximating (indicator) functions $\mathscr{R}$, regularization parameter λ and roughness penalty terms are defined as in Cinquemani (2019). The statistics of the estimation results are reported in Fig. 1 and Fig. 2, for a fixed, educated choice of $\lambda = 10^6$ as well as for λ chosen from the data in every estimation run. The whole simulation was implemented and tested in Matlab on a 3GHz Intel Xeon (Ubuntu) laptop, using standard ODE solvers and CVX (CVX Research, Inc., 2012) for the solution of the optimization problems.

The first observation is that the prodecure is numerically viable. The computation of the function transforms $\mathscr{V}$ takes about 3 minutes, while the solution of every optimization problem (with $N = 156$ unknown parameters, as many as the size of $\mathscr{R}$) takes about 3 seconds (for the automated choice of λ, optimization is repeated several times in the search of its best value). From the statistics of the estimates $\hat{\rho}_E$ in Fig. 1, the second observation is that estimates for both the cases $G = G'$ and $G = G''$ have limited bias (no bias at most lags δ), with increasing uncertainty toward the tail of $\bar{\rho}_E$. Here, the automated choice of the regularization parameter λ reduces estimation error relative to a fixed λ. The increasing error in the tails has to be understood as the result of the memory of the reaction network vanishing for large lag times δ, and the smaller amount of measurements effectively entering the computation of $\hat{\rho}_E$ at large lags. All this is in perfect analogy with the results in absence of a modulat-

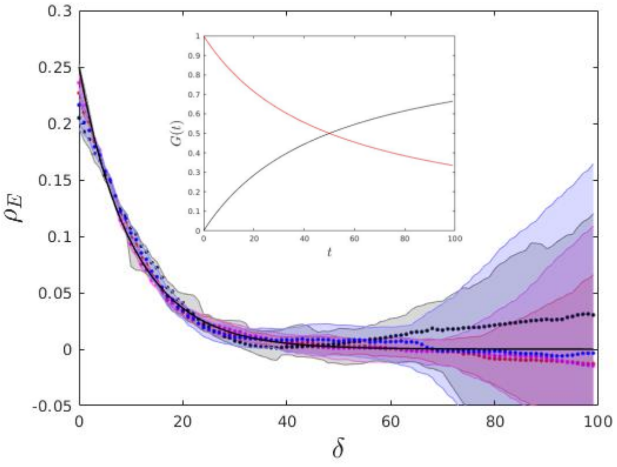

Fig. 1. Results from 100 runs of the estimation of $\bar{\rho}_E$. Solid black line: True profile of $\bar{\rho}_E$; Dotted lines and shaded regions: At all values of δ, median of 100 estimates $\hat{\rho}_E(\delta)$ and region between the 10% and 90% quantiles of the estimates, for fixed ($\lambda = 10^6$, blue and magenta) and automated (grey and red) choices of the regularization factor, in the case $G = G'$ (magenta and red) and $G = G''$ (blue and grey). Plots are bottom-cropped for better scaling. The inset shows the modulating functions G' (red) and G'' (black).

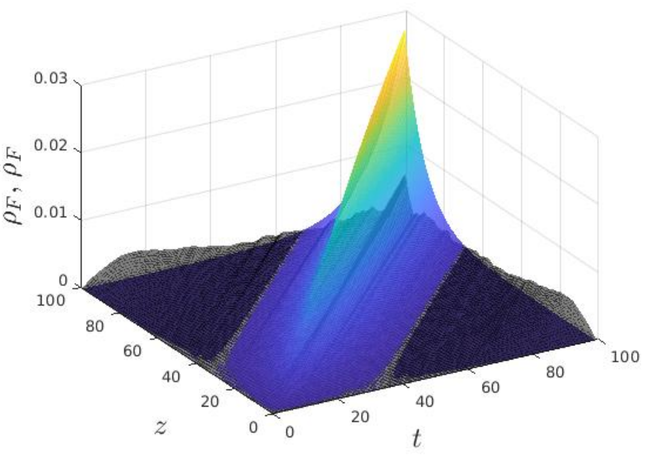

Fig. 2. Statistics of the estimation error $\tilde{\rho}_F(z,t) = \hat{\rho}_F(z,t) - \rho_F(z,t)$ (case $G = G''$, automated choice of λ). At every point (z,t), true autocovariance function $\rho_F(z,t)$ (colored shading) vs. maximal (absolute) value among the 10% and 90% quantile of $\tilde{\rho}_F(z,t)$ over 100 simulation runs (grey mesh).

ing signal of Cinquemani (2019). Interestingly, estimation performance is clearly worse for G'' than for G' in the tails of $\bar{\rho}_E$ (larger estimation error variability) and also around 0 (larger bias). In our interpretation, this is strictly related with the null value of G'' at time zero, which reduces the sensitivity of the data to the information-rich (transient) system dynamics at early times. On the other hand, the estimation error of $\hat{\rho}_E$ does not equally reflect in the estimation error of $\hat{\rho}_F$. Indeed, comparing Fig. 1 with Fig. 2, one finds that the error strengh in the estimation of ρ_F is smaller (in a relative sense) than for $\hat{\rho}_E$ for large differences $|z - t|$. Thus, in general, performance should be assessed based on the context, depending whether the input $(E(t))$ or the controlled $(G(t)E(t))$ process is the object of primary interest. In summary, the procedure is viable and capable to estimate the statistics of a modulated noise process, with performance depending on the modulating (control) signal.

6. CONCLUSIONS

Starting from own previous work, in this paper we have developed an approach for the effective numerical solution of so-called GMEs for a class of modulated input processes where the modulating input is in the form of (or can be approximated by) splines. We have used this result to develop an algorithm for the estimation of the statistics of a controlled promoter activation process from gene expression population-snapshot data. Demonstrated via simulation, the method successfully reconstructed the promoter activity autocovariance function in the limits of a realistic dataset. The solution showed interesting dependencies on the choice of the control input that deserve future quantitative investigation. While developed with specific reference to gene expression reporter systems, many results of this paper are applicable to more general reaction networks and in other fields, where spline approximation of control signals and analysis of controlled continuous-time Markov chains is of relevance.

REFERENCES

Berthoumieux, S., de Jong, H., Baptist, G., Pinel, C., Ranquet, C., Ropers, D., and Geiselmann, J. (2013). Shared control of gene expression in bacteria by transcription factors and global physiology of the cell. *Molecular Systems Biology*, 9(1), 634.

Chait, R., Ruess, J., Bergmiller, T., Tkacik, G., and Guet, C. (2017). Shaping bacterial population behavior through computer-interfaced control of individual cells. *Nature Communications*, 8, 1535.

Cinquemani, E. (2019). Stochastic reaction networks with input processes: Analysis and application to gene expression inference. *Automatica*, 150–156.

CVX Research, Inc. (2012). CVX: Matlab software for disciplined convex programming. `http://cvxr.com/cvx`.

de Jong, H., Ranquet, C., Ropers, D., Pinel, C., and Geiselmann, J. (2010). Experimental and computational validation of models of fluorescent and luminescent reporter genes in bacteria. *BMC Syst. Biol.*, 4(1), 55.

De Nicolao, G., Sparacino, G., and Cobelli, C. (1997). Nonparametric input estimation in physiological systems: Problems, methods, and case studies. *Automatica*, 33(5), 851 – 870.

Fiore, G., Perrino, G., di Bernardo, M., and di Bernardo, D. (2016). In vivo real-time control of gene expression: A comparative analysis of feedback control strategies in yeast. *ACS Synth Biol*, 5, 154–62.

Gillespie, D. (1992). A rigorous derivation of the chemical master equation. *Physica A*, 188, 404–425.

Hasenauer, J., Waldherr, S., Doszczak, M., Radde, N., Scheurich, P., and Allgower, F. (2011). Identification of models of heterogeneous cell populations from population snapshot data. *BMC Bioinformatics*, 12(1), 125.

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

Komorowski, M., Finkenstädt, B., Harper, C., and Rand, D. (2009). Bayesian inference of biochemical kinetic parameters using the linear noise approximation. *BMC Bioinformatics*, 10(1), 343.

Lestas, I., Paulsson, J., Ross, N.E., and Vinnicombe, G. (2008). Noise in gene regulatory networks. *IEEE Trans. Autom. Control.*, 53(Special Issue), 189–200.

Milias-Argeitis, A., Rullan, M., Aoki, S.K., Buchmann, P., and Khammash, M. (2016). Automated optogenetic feedback control for precise and robust regulation of gene expression and cell growth. *Nature Communications*, 7, 12546.

Munsky, B., Trinh, B., and Khammash, M. (2009). Listening to the noise: Random fluctuations reveal gene network parameters. *Mol. Syst. Biol.*, 5(318).

Neuert, G., Munsky, B., Tan, R., Teytelman, L., Khammash, M., and van Oudenaarden, A. (2013). Systematic identification of signal-activated stochastic gene regulation. *Science*, 339(6119), 584–587.

Ocone, A., Haghverdi, L., Mueller, N., and Theis, F. (2015). Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data. *Bioinformatics*, 31(12), i89–i96.

Paulsson, J. (2005). Models of stochastic gene expression. *Phys. Life Rev.*, 2(2), 157 – 175.

Raj, A. and van Oudenaarden, A. (2008). Nature, nurture, or chance: Stochastic gene expression and its consequences. *Cell*, 135, 216–226.

Rao, C., Wolf, D., and Arkin, A. (2002). Control, exploitation and tolerance of intracellular noise. *Nature*, 420, 231–237.

Swain, P., Elowitz, M., and Siggia, E. (2002). Intrinsic and extrinsic contributions to stochasticity in gene expression. *PNAS*, 99(20), 12795–12800.

Thattai, M. and van Oudenaarden, A. (2001). Intrinsic noise in gene regulatory networks. *PNAS*, 98(15), 8614–8619.

Uhlendorf, J., Miermont, A., Delaveau, T., Charvin, G., Fages, F., Bottani, S., Batt, G., and Hersen, P. (2012). Long-term model predictive control of gene expression at the population and single-cell levels. *PNAS*, 109(35), 14271–14276.

Wahba, G. (1990). *Spline models for observational data.* SIAM, Philadelphia, USA.

Zechner, C., Ruess, J., Krenn, P., Pelet, S., Peter, M., Lygeros, J., and Koeppl, H. (2012). Moment-based inference predicts bimodality in transient gene expression. *PNAS*, 109(21), 8340–8345.

Zhao, Z. and Li, X. (2013). Inference for modulated stationary processes. *Bernoulli*, 19(1), 205–227.