Identification of stochastic gene expression models over lineage trees

Aline Marguet, Eugenio Cinquemani

Univ. Grenoble Alpes, Inria, 38000 Grenoble, France
(e-mail: aline.marguet@inria.fr, eugenio.cinquemani@inria.fr)

Abstract: In previous work, we have developed an autoregressive Mixed-Effects model of the evolution of the kinetic gene expression parameters along cell generations, and an identification method simultaneously exploiting single-cell gene expression profiles and known parental relationships among cells (lineage tree data). Here, we extend our modelling and identification approach to explicitly account for stochasticity of promoter activation, and demonstrate via simulation the performance of the method and the improvement relative to the original approach where this source of noise is not accounted for.

Keywords: Branching process, Monte Carlo optimization, Extrinsic noise, HMM, Filtering

1. INTRODUCTION

Modern experimental technologies for the monitoring of single-cell dynamics have unveiled variability of gene expression across cells of a genetically identical population (Elowitz, 2002). Quantitative characterization of gene expression response variability has become an important challenge for the understanding of cellular mechanisms that reduce noise or exploit stochasticity for survival and diversification (Raj and van Oudenaarden, 2008). Among the sources of gene expression noise is the variability of cellular physiology across cells or in the lifetime of every individual cell. Referred to as extrinsic noise (Swain et al., 2002), this is often believed to dominate the so-called intrinsic noise, which instead refers to randomness of the transcription and translation events that determine the expression of a gene (Llamosi et al., 2016). Most approaches to the modelling and identification of gene expression noise from single-cell data assume that individual cells are statistically independent of each other (Munsky et al., 2009; Zechner et al., 2014). However, models of cell division in growing populations as well as experimental results show that correlations may play an important role in shaping variability through the population (Thomas, 2017; Ferraro et al., 2016; Taheri-Araghi et al., 2015).

Focusing on cell-to-cell variability, extrinsic noise can be described in terms of individual kinetic gene expression parameters taking random values from a common population distribution, *i.e.* by a Mixed-Effects (ME) modelling framework (Llamosi et al., 2016). To further account for correlation among individuals, in Marguet et al. (2019), we proposed an extension of ME modelling and identification called ARME. Here, inheritance and variability of individual cell parameters at division is described via an Auto-Regressive process (whence the prefix AR), and known parental relationships among the observed cells are explicitly taken into account. Simulations as well as application to real data showed that ARME improves reconstruction of statistical parameter variability across cells, and outperforms state-of-the-art indirect methods for the reconstruction of correlations among individuals. Yet, variability of kinetic parameters along the lifetime of a cell is not taken into account.

In this paper, we further develop ARME so as to cope with extrinsic gene expression variability within individual cells. We focus on the key phenomenon of random single-cell promoter activation in response to a common environmental stimulus (Suter et al., 2011). Taken cell state and parameter inheritance into account, the model we propose is a linear continuous-time stochastic dynamical model defined over the lineage tree, with jumps at cell division. Based on this, we pose identification of the inheritance kernel and other population parameters as a maximum likelihood problem, which is then reconducted to a filtering problem over trees (Chou et al., 1994; Desbouvries et al., 2006; Durand et al., 2004). A numerical approach for the solution of this problem is developed as a novel extension of the randomized approach proposed in Marguet et al. (2019). By means of simulations in a variety of scenarios, we show that the method provides unbiased estimates of the population parameters sought, and also demonstrate how failing to account for promoter response noise yields biased estimates of the same parameters.

The paper unfolds as follows. In Sec. 2 we review ARME modelling and identification from Marguet et al. (2019). In Sec. 3 we develop the model and identification method for stochastic promoter response. Simulation results are discussed in Sec. 4, and conclusions are drawn in Sec. 5.

2. ARME MODELLING AND IDENTIFICATION OF GENE EXPRESSION DYNAMICS

In this section we review the modelling framework and the identification approach that we developed in Marguet et al. (2019).

⋆ 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.

2.1 *Modelling gene expression kinetics over lineage trees*

Let V be a set of indices for individual cells. The expression dynamics of a gene of interest in a cell $v \in V$ is described by the model

$$\begin{cases} \dot{m}^v(t) \;=\; k_m^v u(t) - g_m^v m^v(t), \\ \dot{p}^v(t) \;\;=\; k_p^v m^v(t) - g_p^v p^v(t), \end{cases} \qquad t \geq t_0^v, \qquad (1)$$

where t is a universal time reference, and t_0^v is the initial time of cell v. In this model, $u(t)$ quantifies the strength of promoter activation in response of a common environmental stimulus $s(t)$, $m^v(t)$ the intracellular concentration of mRNA molecules transcribed at a rate k_m^v, $p^v(t)$ the concentration of protein molecules translated from mRNAs at a rate k_p^v. Rates g_m^v and g_p^v incorporate molecular degradation and growth-related dilution for mRNAs and proteins. Let $s(t)$ take value 1 when the stimulus is present and 0 othewise. In Marguet et al. (2019), similar to earlier work (Munsky et al., 2009; Llamosi et al., 2016), we considered scenarios where $s(t)$ is known (as in control experiments) and $u(t) = \mathscr{R}\big(s(t)\big)$, with $\mathscr{R}$ a known response functional (in the simplest case, $u(t) = s(t)$). Since $u(t)$ is determined by the knowledge of s and $\mathscr{R}$, for the rest of this section we may neglect $s(t)$ and simply refer to $u(t)$ (this will not be the case for the stochastic model of Sec. 3.1). Note that $u(t)$ is assumed identical over all cells $v \in V$. Due to its deterministic nature, the model is best suited to genes with limited expression noise.

Let $\psi^v = (k_m^v, g_m^v, k_p^v, g_p^v)$ denote the d-dimensional vector of individual cell parameters ($d = 4$). Variability of gene expression across cells is captured in part by different values of ψ^v across cells V. By standard ME modelling (Llamosi et al., 2016), $\{\psi^v\}_{v\in V}$ are described as identically distributed independent (i.i.d.) random variables with a common population distribution. In our ARME framework instead, we introduce correlation among cells in terms of stochastic inheritance of parameters from mother to daughter cells. Let us see V as nodes of a graph and let $W \subseteq V \times V$ be a tree. Biologically, $(v^-, v) \in W$ denotes cells in a parental relationship, with v^- the direct ancestor (mother) of v. We assume that the ψ^v obey the AR model

$$\varphi^v = A\varphi^{v^-} + (I - A)b + \eta^v, \qquad \psi^v = \exp(\varphi^v), \qquad (2)$$

where $A \in \mathbb{R}^{d\times d}$, $b \in \mathbb{R}^d$, I denotes the identity matrix, and the η^v are size-d i.i.d. Gaussian random vectors, $\eta^v \sim \mathscr{N}(0, \Omega)$, also independent of φ^{v^-}. The componentwise exponential transformation makes the entries of ψ^v log-normally distributed, thus nonnegative. We further assume that A is strictly stable and that the process φ^v is in a weakly stationary regime, in the sense that one can define a common population mean $\mathbb{E}[\varphi^v] = \mu$ and covariance matrix $\mathrm{Var}(\varphi^v) = \Sigma$. Then $\mu = b$ and Σ obeys $\Sigma = A\Sigma A^T + \Omega$. For different values of A, this model expresses the extent to which the offspring parameters φ^v are determined by (inherited from) the parent parameters φ^{v^-}. Indeed $\mathrm{Cov}(\varphi^v, \varphi^{v^-}) = A\Sigma$, *i.e.* A is the (matrix) correlation coefficient between φ^v and φ^{v^-}. In particular, for A diagonal with $A_{i,i} \in [0, 1)$, $i = 0, \ldots, d$, inheritance applies separately to every parameter, the closer the $A_{i,i}$ to 0 (resp. 1), the smaller (resp. larger) the degree of inheritance. For $A = 0$ (statistically independent individuals), the model reduces to standard ME, with common population distribution $\varphi^v = \log\psi^v \sim \mathscr{N}(b, \Omega)$.

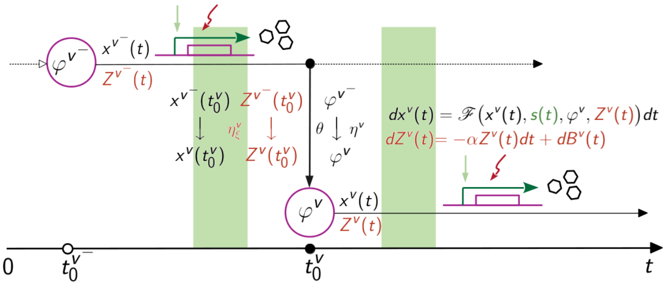

Fig. 1. Cell division and gene expression dynamics. Green: Stimuli; Red: Stochastic extension of the model.

Depending on the organism, one or several offspring v may correspond to the same mother cell v^-, corresponding to different graphs W. In Marguet et al. (2019), we considered bacterial mytosis, *i.e.* a mother cell splits into two cells at a common time t_0^v, and yeast budding, *i.e.* several cells v generated at different times t_0^v can be daughters of the same mother v^-. We assumed that mRNA and protein molecule concentrations are carried over from mother to daughter cells at division. That is, denoting the state vector of cell v as $x^v(t) = [m^v(t), p^v(t)]^T$,

$$x^v(t_0^v) = x^{v^-}(t_0^v). \qquad (3)$$

In the sequel, for simplicity, we focus on the case of mytosis and assume that the initial state $x^v(t_0^v)$ for the root $v = 0$ of W is given. Generalizations are immediate (Marguet et al., 2019). The model is illustrated in Fig. 1.

2.2 *Identification using lineage tree data*

In Marguet et al. (2019), we further addressed the problem of identifying the model of the previous section from single-cell gene expression data, with the additional knowledge of the parental relationships W. Let t_j^v, with $j = 1, \ldots, n^v$, be n^v measurement times for cell $v \in V$. We assume that single-cell gene expression measurements obey the model

$$y_j^v = Cx^v(t_j^v) + h\varepsilon_j^v, \qquad j = 1, \ldots, n^v, \qquad (4)$$

where $x^v(t_j^v)$ is the cell state evolving in accordance with dynamics (1), terms ε_j^v are zero-mean unit-variance Gaussian random variables independent across j and v, and C is known. We restrict attention to the case where $C = [0, 1]$ (the observed variable is the protein concentration p), so that ε_j^v is scalar, and $h > 0$ is an unknown constant that defines the measurement error strength.

Let $\theta = (A, b, \Omega, h)$, $Y^v = \{y_j^v : j = 1, \ldots, n^v\}$ and $Y = \{Y^v : v \in V\}$. The identification problem is to estimate θ given Y and W. The solution developed rests on maximum likelihood, that is, we seek the estimator

$$\hat{\theta}(Y, W) = \arg\max_{\theta\in\Theta} L(\theta|Y, W)$$

(assuming the maximum exists), where $L(\theta|Y, W)$ is the log-likelihood $\log p(Y|W, \theta)$ and Θ is a suitable parameter space. In a standard ME context, where $W = \emptyset$ (parental relationships are not considered) and $A = 0$ (statistically independent individuals), an effective numerical approach to estimate (b, Ω, h) is the randomized optimization method SAEM (Stochastic Approximation of Expectation Maximization (Delyon et al., 1999)). In Marguet et al. (2019), we extended SAEM into our ARME identification method, which exploits the additional knowledge of W and estimates A as well. We now briefly describe SAEM and its extension into ARME identification.

Let $\varphi = \{\varphi^v : v \in V\}$ be all individual cell parameters. Both methods generate a (probabilistically convergent) sequence of estimates $\hat{\theta}_k$ by the following iteration on k:

- S-step: simulate φ_{k+1} according to $p(\varphi|Y, W, \hat{\theta}_k)$;
- E-step: compute
  $Q_{k+1}(\theta) = Q_k(\theta) + \lambda_k(\log(p(Y, \varphi_{k+1}|W, \theta)) - Q_k(\theta))$;
- M-step: update $\hat{\theta}_{k+1} = \arg\max_\theta Q_{k+1}(\theta)$,

where $\lambda_k > 0$ is a suitable annealing sequence. The critical step is the S-step, where a sample value φ_{k+1} of φ needs to be randomly generated in accordance with $p(\varphi|Y, W, \hat{\theta}_k)$, the conditional distribution of φ given the current estimate $\hat{\theta}_k$ of θ. The S-step can be implemented via Metropolis-Hastings. Writing θ in place of $\hat{\theta}_k$ for simplicity, at every iteration k, this amounts to simulate a Markov chain φ_j, using a proposal distribution $q(\tilde{\varphi}_{j+1}, \varphi_j)$, and accepting $\tilde{\varphi}_{j+1}$ as the new state of the chain with probability

$$\min\left\{1, \frac{p(\tilde{\varphi}_{j+1}|Y, W, \theta)}{p(\varphi_j|Y, W, \theta)} \frac{q(\tilde{\varphi}_{j+1}, \varphi_j)}{q(\varphi_j, \tilde{\varphi}_{j+1})}\right\}.$$

For SAEM, in view of $W = \emptyset$ and statistical independence of the individuals, the problem can be separated out into $|V|$ simpler simulation problems, one per individual. Mathematically, this follows from the fact that term

$$p(\varphi|Y, W, \theta) \propto p(Y|\varphi, W, h)p(\varphi|W, \theta) \tag{5}$$

can be factorized as $\prod_{v\in V} p(Y^v|\varphi^v, h)p(\varphi^v|\theta)$. In ARME instead, Metropolis-Hastings needs to be performed at once on the whole tree W. Indeed $p(\varphi|W, \theta)$ is determined by the autoregression (2) and it cannot be factored out into individual cell terms. This largely complicates the choice of the proposal distribution q for achieving suitable acceptance rates and practical convergence. In Marguet et al. (2019), an effective sampling strategy was developed that is based on a hierarchy of proposal distributions incorporating whole-tree, generation, and individual-cell level sampling. In addition to the convergence of the method, we showed via numerical simulation that estimation of A in particular, and of all entries of θ in general, is significantly improved relative to a state-of-the-art competing approach. The reader is referred to the original paper for more details, numerical performance analysis, implementation code and application to real data.

Importantly, in view of the model of Sec. 2.1, the evaluation of the likelihood $p(Y|\varphi, W, h)$ in (5) is simple. For the putative cell parameters φ and any $v \in V$, let $X^v_\varphi = \{x^v_\varphi(t^v_j) : j = 1, \ldots, n^v\}$ be the values of the state of cell v at the measurement times t^v_j. For all v, X^v_φ is determined by φ, since it follows from the solution of the ODE system (1) along the branches of W, using (3) for parent-offspring transitions. Then, in view of (4),

$$\begin{aligned} p(Y|\varphi, W, h) &= \prod_{v\in V} p(Y^v|X^v_\varphi, h) \\ &= \prod_{v\in V}\prod_{j=0}^{n^v} f_h\left(y^v_j - Cx^v_\varphi(t^v_j)\right) \end{aligned} \tag{6}$$

where $f_h(\cdot)$ is the density function of $\mathcal{N}(0, h)$. As we will see, the problem becomes more complicated in presence of noisy dynamics.

3. MODELLING AND IDENTIFICATION IN PRESENCE OF PROMOTER NOISE

Model (1) describes gene expression in terms of deterministic dynamics entirely defined by the individual cell parameters ψ^v. Yet, in general, single-cell response is notoriously noisy (Elowitz, 2002). Among the various sources of noise, in this work we focus on the variability of promoter activation in response to an external stimulus. Our objective is to investigate the importance to account for this source of variability in the identification of the gene expression models from the expression measurements Y and lineage data W. In particular, we are interested in the accuracy of reconstruction of the parameter inheritance dynamics (2). To address this question, a simple extension of model (1) is proposed in Sec. 3.1, and a corresponding extension of the ARME identification approach is discussed in Sec. 3.2. Performance comparison between this extended identification method and the original method of Sec. 2 will be developed in Sec. 4.

3.1 Stochastic promoter activation model

For a given cell v, let us interpret $u(t)$ in (1) as the result of a random response to the stimulus $s(t)$. We model this as follows. For $i \in \{0, 1\}$, let $\mu_i \in \mathbb{R}_+$ and $\gamma_i \in \mathbb{R}_+$. Let $Z^v(t)$ be a stationary Ornstein-Uhlenbeck process described by $dZ^v(t) = -\alpha Z^v(t)dt + dB^v(t)$, with B^v standard Brownian motion and $\alpha > 0$. We let u be a random outcome of the switching process

$$U^v(t) = \mu_{s(t)} + \gamma_{s(t)} Z^v(t) \tag{7}$$

with $(\mu_{s(t)}, \gamma_{s(t)})$ equal to (μ_0, γ_0) if $s(t) = 0$ and to (μ_1, γ_1) if $s(t) = 1$. We further assume that processes B^v are independent across v. This model describes single-cell promoter activation as a variable response to the common environmental stimulus $s(t)$. Depending on absence or presence of the stimulus, promoter activation fluctuates around μ_0 or μ_1, (with standard deviation proportional to γ_0 and γ_1, respectively) and is different across cells. For the purpose of our study, this model provides a convenient, minimal description of stochastic single-cell response with nontrivial dynamics. The outcomes of U^v are piecewise continuous w.p.1. The normalized autocorrelation function of $U^v(t)$ is $\rho(\tau) = \mathbb{E}[Z^v(t)Z^v(t+\tau)] = \exp(-\alpha|\tau|)/(2\alpha)$, i.e. α defines the "memory" of promoter activation. For $\gamma_0 = \gamma_1 = 0$, a deterministic relationship between $u(t)$ and $s(t)$ of the type of Sec. 2.1 is recovered as a special case. For $\gamma_0 > 0$ or $\gamma_1 > 0$, the interpretation of u as a function of s holds in a stochastic sense, that is, different continuous response profiles u may correspond to a same stimulus s.

In summary, for any $v \in V$ let $\xi^v(t) = [Z^v(t)^T, x^v(t)]^T = [Z^v(t), m^v(t), p^v(t)]^T$. Our gene expression model with stochastic promoter response is given by the stochastic differential equation

$$d\xi^v(t) = F^v\big(s(t)\big)\xi^v(t)dt + f^v\big(s(t)\big)dt + GdB^v(t), \tag{8}$$

with $t \geq t^v_0$, where $G = [1, 0, 0]^T$ and

$$F^v\big(s(t)\big) = \begin{bmatrix} -\alpha & 0 & 0 \\ k^v_m\gamma_{s(t)} & -g^v_m & 0 \\ 0 & k^v_p & -g^v_p \end{bmatrix}, \; f^v\big(s(t)\big) = \begin{bmatrix} 0 \\ k^v_m\mu_{s(t)} \\ 0 \end{bmatrix}.$$

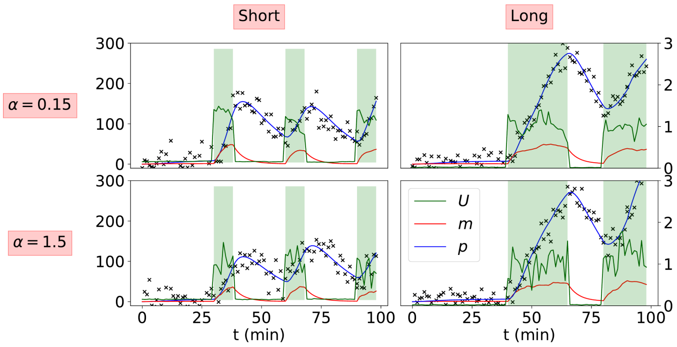

Fig. 2. Example dynamics in one cell v for Short and Long stimuli (green bands: $s(t) = 1$), and different memory α. Green, red, and blue lines: $U^v(t)$, $m(t)$ and $p(t)$; Black crosses: Y^v.

For any $(v^-, v) \in W$, we still assume that (2) holds, and consider a generalization of (3),

$$\xi^v(t_0^v) = A_\xi \xi^{v^-}(t_0^v) + \eta_\xi^v, \qquad (9)$$

with random vectors $\eta_\xi^v \sim \mathscr{N}(0, \Omega_\xi)$ independent across v. The model is illustrated in Fig. 1. Example simulations from this model are in Fig. 2.

3.2 ARME identification with noisy promoter dynamics

In this section we extend ARME identification to stochastic gene expression dynamics. The randomized iteration reviewed in Sec. 2.2 is a general method for likelihood maximization in presence of hidden variables (see Delyon et al. (1999)). In turn, for a suitable proposal distribution q, the Metropolis-Hastings implementation of the M-step is viable as long as (5) can be evaluated. As already noticed, term $p(\varphi|W, \theta)$ is determined by the regression model (2), which is unchanged. On the contrary, because of the stochastic dynamics (8), the derivation of (6) no longer applies. The challenge is therefore the evaluation of $p(Y|\varphi, W, h)$. Note that measurements $y_j^v \in Y$ now obey

$$y_j^v = C_\xi \xi^v(t_j^v) + h\varepsilon_j^v,$$

with $C_\xi = [0, 0, 1]^T$. For simplicity, we next drop φ, W and h from the notation.

Let $v = 0$ be the root of tree W. Let $G_k \subset V$ be the nodes at distance (length of shortest connecting path) k from the root (*i.e.* the cells of generation k), in particular, $G_0 = \{0\}$. Let $\mathscr{P}(v)$ denote the set of nodes in V connecting v to the root node (the "past" of v), and $\mathscr{P}(G_k) = \{G_0 \cup \ldots \cup G_{k-1}\}$. Finally, for any subset S of V, let $Y^S = \{Y^v : v \in S\}$. Assume that the tree has m generations (*i.e.* at least one node $v \in V$ is at distance m from 0 and none is at distance $m + 1$). By the Bayes law, one may write

$$p(Y) = p(Y^{G_0}) \cdot \prod_{k=1}^{m} p(Y^{G_k}|Y^{\mathscr{P}(G_k)}). \qquad (10)$$

This decomposition hints at the evaluation of $p(Y)$ via a Kalman filtering-type recursion (Jazwinski, 1970), where factor $p(Y^{G_k}|Y^{\mathscr{P}(G_k)})$ is determined by the one-step predictor at iteration k. However, the problem is complicated by (i) the variable size of G_k over k (for a binary tree, the size of G_k is 2^{k-1}), and (ii) the stochastic dynamics that relate measurements Y^v of a same cell.

A computational solution to (i) is given by Desbouvries et al. (2006) for a linear Gaussian model over a tree. For a similar scenario, an alternative approach built upon backward filtering from leaves to root is taken by Chou et al. (1994) (and others, see *e.g.* Durand et al. (2004)). Neither of the two approaches addresses point (ii), whereas both fully account for the correlation among measurements along parallel branches of the tree. Correlation across parallel branches is originated by the hidden state dynamics and implies that, at any k, the factorization

$$p(Y^{G_k}|Y^{\mathscr{P}(G_k)}) = \prod_{v \in G_k} p(Y^v|Y^{\mathscr{P}(v)}) \qquad (11)$$

does not hold in general. In the same spirit as Kuzmanovska et al. (2017), we instead make an approximation, and take (11) as our working assumption. That is, we make the hypothesis that the information about Y^v contained in past generations is entirely captured by the direct ancestors of v. Whereas the accuracy of the approximation is a priori unclear, it allows us to greatly simplify the calculation of (10). This is extremely important because this calculation enters every sample of a randomized optimization algorithm. The computation goes as follows.

For any k and $v \in G_k$, with a slight abuse of notation, let $Y_{\mathscr{P}(j)}^v$ denote the past measurements of cell v at time t_j. That is, for $j = 1, \ldots, n_v$, $Y_{\mathscr{P}(j)}^v = \{y_1^v, \ldots, y_{j-1}^v\}$. Then

$$p(Y^v|Y^{\mathscr{P}(v)}) = \prod_{j=1}^{n^v} p(y_j^v|Y_{\mathscr{P}(j)}^v, Y^{\mathscr{P}(v)}). \qquad (12)$$

Since in our assumptions Y and $\{\xi^v\}_{v \in V}$ are jointly Gaussian processes, these conditional probabilities can be evaluated exactly via a Kalman recursion. Let $\hat{\xi}_{j|i}^v$ and $\Pi_{j|i}^v$ be the conditional mean and covariance matrix of $\xi^v(t_j)$ given $Y_{\mathscr{P}(i)}^v$ and $Y^{\mathscr{P}(v)}$. For the relevant indices j from 1 to n^v, it holds that

$$\begin{aligned}
\hat{\xi}_{j|j}^v &= \hat{\xi}_{j|j-1}^v + \Pi_{j|j-1}^v C_\xi^T (\Lambda_j^v)^{-1}(y_j^v - \hat{y}_j^v), \\
\Pi_{j|j}^v &= \Pi_{j|j-1}^v - \Pi_{j|j-1}^v C_\xi^T (\Lambda_j^v)^{-1} C_\xi \Pi_{j|j-1}^v, \\
\hat{\xi}_{j+1|j}^v &= \Phi^v(t_{j+1}^v, t_j^v, \hat{\xi}_{j|j}^v), \\
\Pi_{j+1|j}^v &= \Psi^v(t_{j+1}^v, t_j^v, \Pi_{j|j}^v),
\end{aligned} \qquad (13)$$

where $\hat{y}_j^v = C_\xi \hat{\xi}_{j|j-1}^v$ and $\Lambda_j^v = C_\xi \Pi_{j|j-1}^v C_\xi^T + h^2$ are, respectively, the conditional mean and covariance matrix of y_j^v given $Y_{\mathscr{P}(j)}^v$ and $Y^{\mathscr{P}(v)}$ (Jazwinski, 1970). In turn, for generic $\bar{\xi}$, $\bar{\Pi}$, τ and t, $\Phi^v(\tau, t, \bar{\xi})$ and $\Psi^v(\tau, t, \bar{\Pi})$ are the solution at τ of

$$\begin{aligned}
&\dot{\xi}(\cdot) = F^v\big(s(\cdot)\big)\xi(\cdot) + f^v\big(s(\cdot)\big), && \xi(t) = \bar{\xi}, \\
&\dot{\Pi}(\cdot) = F^v\big(s(\cdot)\big)\Pi(\cdot) + \Pi(\cdot)F^v\big(s(\cdot)\big)^T + GG^T, && \Pi(t) = \bar{\Pi}.
\end{aligned}$$

The recursion is initialized with

$$\hat{\xi}_{1|0}^v = \Phi^v(t_1^v, t_0^v, \hat{\xi}_0^v), \quad \Pi_{1|0}^v = \Psi^v(t_1^v, t_0^v, \Pi_0^v), \qquad (14)$$

where $\hat{\xi}_0^v$ and Π_0^v are the state estimate for cell v at initial time t_0^v. If v has a known parent v^-, these are inherited from the latest estimate of the parent state ξ^{v^-} before t_0^v, in the light of (9). For $J = \max\{j : t_j^{v^-} \le t_0^v\}$, one gets

$$\begin{aligned}
\hat{\xi}_0^v &= A_\xi \Phi^{v^-}(t_0^v, t_J^{v^-}, \hat{\xi}_{J|J}^{v^-}), \\
\Pi_0^v &= A_\xi \Psi^{v^-}(t_0^v, t_J^{v^-}, \Pi_{J|J}^{v^-}) A_\xi^T + \Omega_\xi.
\end{aligned} \qquad (15)$$

Otherwise, if $v = 0$ is the root note of W, $\hat{\xi}_0^v$ and Π_0^v simply express a Bayes prior on the initial state. Finally, leveraging this iterative calculation of quantities $\hat{y}_j^v$ and Λ_j^v, the factors in (12) may be evaluated as

$$p(y_j^v | Y_{\mathscr{P}(j)}^v, Y^{\mathscr{P}(v)}) = f_{\Lambda_j^v}(y_j^v - \hat{y}_j^v) \qquad (16)$$

(recall that $f_{\Lambda_j^v}$ stands for density function of $\mathscr{N}(0, \Lambda_j^v)$).

In summary, the proposed identification algorithm for the model of Sec. 3.1 is analogous to the SAEM algorithm of Sec. 2.2, with the crucial difference that, for putative parameter values φ and h, the likelihood expression (6) is replaced by the calculation of (10) under approximation (11). The factors of (11) are calculated iteratively by propagating the filtering equations (13) and (14)–(15) from root to leaves of W, evaluating the factors (16) alongside the filtering iterations, and forming the products (12). Despite the approximation (11), we show in the next section that the proposed method significantly outperforms the method of Sec. 2.2 in the case of data from stochastic promoter activation. Development of an exact algorithm and comparison of performance with the current appoximation is instead left for future studies.

4. SIMULATION STUDY

We now discuss performance of identification of the population parameters $\theta = (A, b, \Omega, h)$ in the case of noisy promoter activation. To do this we rely on synthetic datasets generated in accordance with the stochastic promoter model of Sec. 3.1, and run the identification method developed in Sec. 3.2. To evaluate the importance of explicitly accounting for stochastic single-cell dynamics in the identification of θ, on the same datasets, we also run the original method of Sec. 2.2, which assumes identical promoter activation for all cells, and compare results. In the sequel, we refer to the two identification algorithms as SI (Stochastic model-based Identification) and DI (Deterministic model-based Identification), respectively. All algorithms are implemented in Julia.

We consider cells undergoing mitosis, *i.e.* every cell eventually splits into two daughter cells. We consider two types of datasets corresponding to different structures W of the observed population. The first type (which we refer to as "Tree"), is a binary tree spanning 7 generations (both daughters of every mother cell are kept track of), for a total of $2^7-1 = 127$ cells. The second type ("Branch") is a single branch spanning 127 generations (only one daughter of every mother cell is kept track of), for the same total of 127 cells. The two types correspond to different experimental videomicroscopy scenarios (cells in a microfluidic chamber, as *e.g.* in Elowitz (2002), and in a "mother machine" as *e.g.* in Taheri-Araghi et al. (2015)). Notice that, in the latter case, the identification method of Sec. 3.2 is exact since (11) holds without any approximation.

We express time t in minutes. We simulate a population with cell divisions every 90 minutes and parameters θ fixed as in Marguet et al. (2019). In particular, $A = 0.5I$, $\exp(b) = [0.294, 0.947, 0.1, 10]^T$, Ω is non-diagonal and $h = 20$ (rate units are minutes^{-1}, concentrations are in arbitrary units). Measurements are taken every minute. We assume noiseless state inheritance (*i.e.* $A_\xi = 0$ and $\Omega_\xi = 0$). For promoter activation we consider four different scenarios. We consider external perturbation signals $s(t)$ with a "Short" and "Long" duration of stimuli (8-minute stimuli every 30 minutes, and 25-minute stimuli every 40 minutes, respectively). For both Short and Long stimuli, we consider stochastic response with long memory ($\alpha = 0.15$) and short memory ($\alpha = 1.5$). Stochastic promoter response has small mean ($\mu_0 = 0.02$) in absence of stimulus ($s(t) = 0$), larger mean ($\mu_0 = 1$) in presence of stimulus ($s(t) = 1$). The values of γ_0 and γ_1 are fixed depending on α, such that, for both values of α tested, the (stationary) standard deviation of U^v is equal to 0.006 when $s(t) = 0$ and 0.29 when $s(t) = 1$. Fig. 2 illustrates the four scenarios.

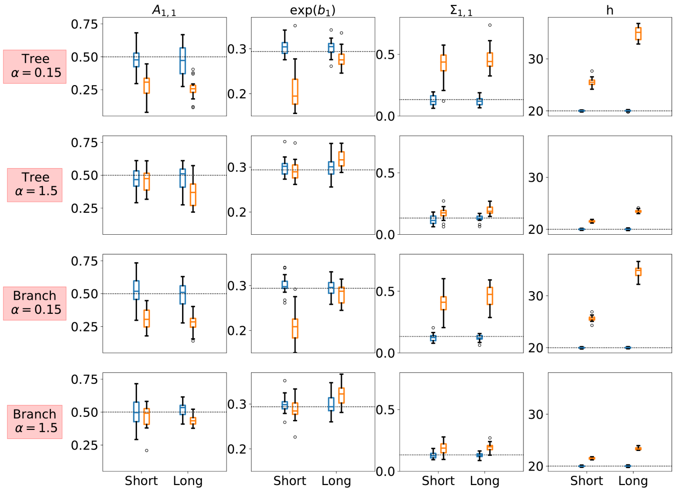

Fig. 3. Identification statistics in different scenarios for the entries of θ pertaining g_m. Blue: SI; Orange: DI.

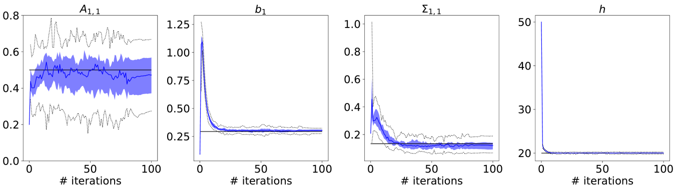

Fig. 4. Statistics of the SI iterations over 20 datasets. Black line: true value; Blue line, band and black envelope correspond to the size of boxes and whiskers in Fig. 3

For each of the four scenarios, we simulate 20 datasets and run the two identification algorithms on every dataset (∼1h for a Tree and ∼5h for a Branch on a modern desktop). We assume A diagonal and, to avoid known identifiability issues that would blur the analysis (Marguet et al., 2019), we assume that k_m is given. For the parameters under estimation, we start the identification methods from initial guesses off the truth by at least 100% of the true value. Identification results in the form of statistics over the different simulated datasets are reported in Fig. 3. Convergence of the iterative SI algorithm is shown in Fig. 4. For reasons of space, we only report results for the entries of θ corresponding to parameter g_m. These results are representative of identification performance for the remaining entries of φ (we omit discussion of cross-correlation terms).

The first observation is that the SI estimates are unbiased. This is true even for the Tree scenario, where (11) is indeed an approximation. On the contrary, DI is severely biased especially in the case of long shocks, where more noise is brought about by random promoter activation. In particular, inheritance (term $A_{1,1}$) is underestimated while parameter variability (term $\Sigma_{1,1}$, as determined from the estimated A and Ω) and measurement noise strength (h) are overestimated. Intuitively speaking, DI attributes randomness across single-cell responses to the modelled sources of noise, notably noise at division, which results in reduced ability to detect inheritance.

The second observation concerns the role of the promoter activation memory. For large α (short memory), the performance gap between DI and SI is less pronounced. This can be understood in terms of relative time scales: Fast fluctuations of promoter response relative to the time scales of the m and p dynamics are averaged away along time. Conversely, slow fluctuations may result in a drift of the dynamics of m and p over a whole cell lifespan, with a similar effect as that of a cell having different kinetic parameters φ. Again, this hampers the ability to estimate mean and inheritance statistics of φ unless promoter fluctuations are explicitly modelled.

An additional observation concerns different experiment designs. Performance of both DI and SI appears to be the same in the Tree and the Branch scenarios considered. With the number of monitored cells and mother-daughter couples being the same in the two cases, in agreement with intuition, this hints that the achievable performance is essentially independent from the number of generations observed. Thus, for equivalent popoulation size, the most convenient experimental setup should be chosen based on other considerations (duration, technical feasibility, probability of mutations, etc.). Also, this observation suggests that the approximation that SI rests upon for a Tree has a rather marginal impact on performance.

In summary, the lack of modelling of random promoter response in DI leads to identification bias, notably inheritance underestimation, especially for long-memory promoter fluctuations and long periods of stochastic response. Unbiased estimates and better performance overall are instead obtained with SI, even under the approximate assumption (11), thanks to the modelling of promoter noise.

5. CONCLUSION

Starting from previous work on modelling and identification of inheritance dynamics for kinetic gene expression parameters, in this work we extended modelling so as to include random single-cell response to external stimuli, and developed the identification approach to take this noise source into account. We showed via simulation the performance of the method developed and its ability to compensate for identification biases that appear if randomness in single-cell response is not accounted for at the identification stage. Future directions of research include mathematical analysis of the method, removal of a conditional independence approximation, application to other dynamical systems with tree-structured dependencies and to real data.

REFERENCES

Chou, K., Willsky, A., and Benveniste, A. (1994). Multiscale recursive estimation, data fusion, and regularization. *IEEE Transactions on Automatic Control*, 39(3), 464–478.

Delyon, B., Lavielle, M., and Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist.*, 27(1), 94–128.

Desbouvries, F., Lecomte, J., and Pieczynski, W. (2006). Kalman filtering in pairwise Markov trees. *Signal Processing*, 86(5), 1049–1054.

Durand, J.B., Gonalvs, P., and Gudon, Y. (2004). Computational Methods for Hidden Markov Tree Models – An Application to Wavelet Trees. *IEEE Transactions on Signal Processing*, 52(9), 10.

Elowitz, M.B. (2002). Stochastic Gene Expression in a Single Cell. *Science*, 297(5584), 1183–1186.

Ferraro, T., Esposito, E., Mancini, L., Ng, S., Lucas, T., Coppey, M., Dostatni, N., Walczak, A.M., Levine, M., and Lagha, M. (2016). Transcriptional memory in the *drosophila* embryo. *Curr. Biol.*, 26, 212–218.

Jazwinski, A.H. (1970). *Stochastic processes and filtering theory*. Elsevier.

Kuzmanovska, I., Milias-Argeitis, A., Mikelson, J., Zechner, C., and Khammash, M. (2017). Parameter inference for stochastic single-cell dynamics from lineage tree data. *BMC Systems Biology*, 11(1), 52.

Llamosi, A., Gonzalez-Vargas, A.M., Versari, C., Cinquemani, E., Ferrari-Trecate, G., Hersen, P., and Batt, G. (2016). What population reveals about individual cell identity: Single-cell parameter estimation of models of gene expression in yeast. *PLOS Comput. Biol.*, 12(2).

Marguet, A., Lavielle, M., and Cinquemani, E. (2019). Inheritance and variability of kinetic gene expression parameters in microbial cells: modeling and inference from lineage tree data. *Bioinformatics*, 35(14), i586–i595.

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

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

Suter, D.M., Molina, N., Gatfield, D., Schneider, K., Schibler, U., and Naef, F. (2011). Mammalian genes are transcribed with widely different bursting kinetics. *Science*, 332, 472–474.

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

Taheri-Araghi, S., Bradde, S., Sauls, J.T., Hill, N.S., Levin, P.A., Paulsson, J., Vergassola, M., and Jun, S. (2015). Cell-size control and homeostasis in bacteria. *Curr. Biol.*, 25(3), 385 – 391.

Thomas, P. (2017). Making sense of snapshot data: ergodic principle for clonal cell populations. *J. Royal Soc. Interface*, 14(136), 20170467.

Zechner, C., Unger, M., Pelet, S., Peter, M., and Koeppl, H. (2014). Scalable inference of heterogeneous reaction kinetics from pooled single-cell recordings. *Nature Methods*, 11, 197–202.