2024 European Control Conference (ECC)
June 25-28, 2024. Stockholm, Sweden

Inference of tree-structured auto-regressive models of gene expression parameters from generation-snapshot data*

Emrys Reginato, Aline Marguet[1] and Eugenio Cinquemani[1,†]

Abstract— **In previous work, we proposed Auto-Regressive (AR) modelling on population trees for the stochastic transmission of individual-cell kinetic gene expression parameters at cell division. We addressed inference of the AR model parameters from individual gene expression profiles in a growing population, under the assumption of known parental relationships. In this paper, we explore the same inference problem in the case where only the generation that cells belong to is known, while parental relationships are unknown. First assuming that individual-cell parameters are measured directly with known degree of uncertainty, we develop a likelihood-based method that is applicable beyond the specific case of gene expression. Then, for data consisting of gene expression profiles, we extend the method into a pipeline for the identification of the AR model parameters via preliminary reconstruction of individual-cell parameters and their uncertainty. Performance of all methods is demonstrated via simulations inspired from real data.**

I. Introduction

Variability of gene expression kinetics and other phenotypic traits is obiquitous in single-cell experiments over isogenic cell populations. It is at the roots of important phenomena such as bet-hedging and adaptation [15]. Several modelling and inference methods have been developed to explore stochasticity within and across individual cells and explain the variability in the data [19], [7]. Most methods simplify the description of population growth by treating cells as independent individuals, thus neglecting parental relationships [20], [21], [14], [9], [1]. This can introduce bias in the analysis and overlook important phenomena observed in the data [3], [17], [5], [18].

In previous work [12], we considered an AR model for the stochastic transmission of kinetic parameters of gene expression from mother to daughter cells in a growing cellular population, thus generalizing Mixed-Effects (ME) modelling to populations of individuals with tree-structured correlations. We further developed a method to infer the AR model parameters from single-cell gene expression time profiles, under the assumption of known parental relationships among the observed cells. The biological relevance of the approach, which allows one to characterize the onset of phenotypic variability over generations, was demonstrated on the subset of fluorescence microscopy gene expression profiles from [11] for which parental relationships are available.

In this paper, based on the same AR modelling framework, we address inference in the case where parental relationships are unknown. Assuming that the generation a cell belongs to is known, we pursue estimation of the AR model parameters from individual-cell observations within generations, which we call generation-snapshot data. The first motivation for the work is that, as [11] witnesses, parental relationships among cells may be unavailable or nontrivial to obtain even from microscopy experiments (unless mother machines are used [16]). The broader motivation is that the study of inference of tree-structured population models from snapshot data, as collected *e.g.* by flow-cytometry [20], [14], [17], is currently limited. Our work is a first step in this direction.

First assuming direct measurements of individual-cell traits of interest, we develop an exact maximum-likelihood method for inference of the AR model parameters from empirical means within generations, and an approximate generalization of the method incorporating empirical variances. Focusing on the case of a full binary tree with a single ancestor, we show by simulations the important role played by correlations across different generations and by the dynamics stemming from the single ancestor on estimation performance. We then extend the method to cope with indirect measurements of single-cell traits, focusing on the case where the traits of interest are kinetic gene expression rates and observations are gene expression profiles. We show that the method is capable of AR model inference on simulations of the experiments of [11]. While focused on cellular populations, results can be of interest for inference of tree-structured models in any application field (multiresolution analysis, phylogeny, image processing, . . .).

In Sec. II, we discuss the modelling framework and the inference problem. In Sec. III, we develop and assess performance of inference methods from direct individual trait measurements. Sec. IV addresses inference from single-cell gene expression profiles. Conclusions and perspectives are in Sec. V. Proofs are not reported in the interest of space.

II. Modelling evolution of individual cell traits in a growing cell population

Borrowing from [12], we consider the following model for the evolution of single-cell traits in a population of dividing cells. Let $\boldsymbol{\varphi}^v$ be a vector of m real parameters quantifying one or several traits of an individual cell v. We assume that traits are constant over the cell lifespan. We let $\boldsymbol{\varphi}^v$ evolve according to the AR model

$$\boldsymbol{\varphi}^v = A\boldsymbol{\varphi}^{v^-} + (I-A)\mathbf{b} + \boldsymbol{\eta}^v, \qquad (1)$$

where v^- is the mother of cell v, A a diagonal matrix of size m with elements between 0 and 1, $\mathbf{b}$ a vector of size

*Work supported in part by project AnaComBa, Equipe-Action Persyval of the Université Grenoble Alpes. All authors are with Université Grenoble Alpes, Inria, 38000, Grenoble, France

[1] These authors contributed equally to the work.

†Corresponding author, eugenio.cinquemani@inria.fr

m, and $\boldsymbol{\eta}^v$ a Gaussian random vector, independent across v, with variance $\Omega \in \mathcal{M}_m(\mathbb{R})$. By this model, daughter cell parameters are the result of a balance between mother cell parameters $\boldsymbol{\varphi}^{v^-}$ and reference parameters $\mathbf{b}$, plus a noise term that reflects randomness of the newborn cell. Matrix A quantifies persistence: The closer A to the identity, the stronger the influence of the mother cell. We refer to A, $\mathbf{b}$ and Ω as the population parameters. Provided a suitable transformation of $\boldsymbol{\varphi}^v$, this model may account for quantitative constraints and/or non-Gaussian distribution of individual cell traits. In [12], by the transformation $\boldsymbol{\phi}^v = \exp \boldsymbol{\varphi}^v$, the model describes kinetic gene expression parameters $\boldsymbol{\phi}^v$ as non-negative, log-normally distributed random variables. Here, we will not restrict ourselves to a specific (set of) trait(s), but we will come back to the case of single-cell gene expression kinetics in Sec. IV. The model takes explicitly into account the lineage cell tree, that is, tree-structured parental relations over subsequent cell generations. From now on, we refer to the case of a complete binary tree, though several results that follow can be generalized.

Let $v = 0$ be the index of the common ancestor cell at generation 0, and let S_n be the set of indices of cells of generation n (thus $v \in S_n \implies v^- \in S_{n-1}$ for $n > 0$). For every $v \in S_n$, with $n = 0, 1, \ldots$, we assume that a noisy version of the cell trait $\boldsymbol{\varphi}^v$,

$$\tilde{\boldsymbol{\varphi}}^v = \boldsymbol{\varphi}^v + \boldsymbol{\epsilon}^v \qquad (2)$$

is available, where $\boldsymbol{\epsilon}^v \sim \mathcal{N}(0, R(n))$. Eq. (2) may equally represent the noisy measurement of a directly observable trait (*e.g.* cell size at birth), or an estimate of individual traits from indirect measurements (*e.g.* kinetic gene expression rates from single-cell gene expression profiles, see Sec. IV). In [12], under stationarity assumptions, we built inference algorithms for A, $\mathbf{b}$ and Ω assuming known parental relationships. Here, instead, we only assume that the generation of a cell is known, without knowledge of parental relationships. That is, the data is the collections $\{\tilde{\boldsymbol{\varphi}}^v : v \in S_n\}$, with $n = 0, 1, \ldots$, which we call generation-snapshot data.

We aim at developing inference methods based on matching the dynamics of model-predicted statistics with measurement statistics calculated within generations. With the generation index in place of a time index, this is resemblant of existing moment-matching procedures for identification of reaction networks (see *e.g.* [20]), with the major difference that these procedures do not account for population lineages. In our case, due to the underlying tree structure, how to define and relate model and data statistics is not obvious. We focus the analysis on statistics up to second order. We consider that the parameters $\boldsymbol{\varphi}^0$ of the common ancestor are fixed, though generally unknown.

First consider the statistics of $\boldsymbol{\varphi}^v$. Since the AR model is the same along all branches of the lineage tree we can define the same mean $\boldsymbol{\mu}(n) = \mathbb{E}(\boldsymbol{\varphi}^v)$ and variance $\Sigma(n) = \mathrm{Var}(\boldsymbol{\varphi}^v)$ for all $v \in S_n$. As for standard AR processes,

$$\boldsymbol{\mu}(n) = A^n \boldsymbol{\varphi}^0 + (I - A^n)\mathbf{b}, \quad \Sigma(n) = \sum_{i=0}^{n-1} A^i \Omega {A^i}^{\mathbf{T}}. \qquad (3)$$

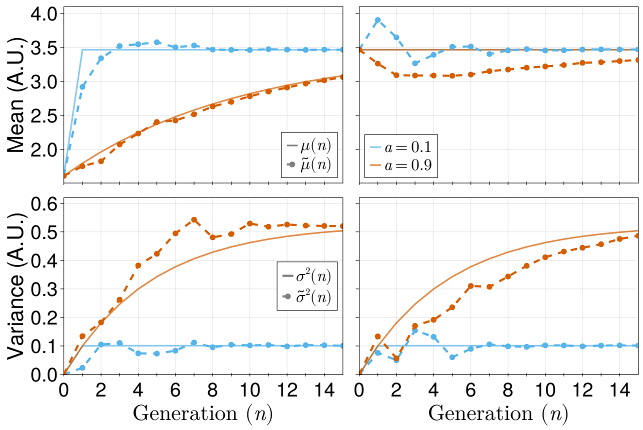

Fig. 1. Moment dynamics in the atypical (left) and typical (right) ancestor scenarios for two values of a. Solid lines: Equations (3); Dash-dotted lines: Empirical moments (4)–(5) from one simulation.

For large n, $\boldsymbol{\mu}(n)$ converges to $\mathbf{b}$ and $\Sigma(n)$ to the solution of $\Sigma = A\Sigma A^{\mathbf{T}} + \Omega$. Next consider the empirical mean and variance of measurements $\tilde{\boldsymbol{\varphi}}^{v_n}$ within different generations,

$$\tilde{\boldsymbol{\mu}}(n) = \frac{1}{|S_n|} \sum_{v \in S_n} \tilde{\boldsymbol{\varphi}}^v \,, \qquad (4)$$

$$\tilde{\Sigma}(n) = \frac{1}{|S_n| - 1} \sum_{v \in S_n} \left(\tilde{\boldsymbol{\varphi}}^v - \tilde{\boldsymbol{\mu}}(n)\right)\left(\tilde{\boldsymbol{\varphi}}^v - \tilde{\boldsymbol{\mu}}(n)\right)^{\mathbf{T}} . \qquad (5)$$

How do (4)–(5) relate with (3) and with the model parameters? Precise relationships will be developed in the next section. Here, to provide an intuition behind the parameter inference methods and results that will follow, we illustrate these relationships by numerical simulations in Julia [2].

Fig. 1 reports example profiles of (3) and (4)–(5) over 16 generations for a scalar model ($m = 1$, with matrices A, Σ, Ω replaced by scalars a, σ^2, ω^2, and non-bold notation for scalars replacing vectors). Empirical statistics are obtained from the random simulation of model (1) with $\omega^2 = 0.1$, and noiseless measurements (2). For the two cases of weak ($a = 0.1$) and strong inheritance ($a = 0.9$), we consider two scenarios: non-stationary ($\varphi^0 = \log(5) \neq b$) and stationary process mean ($\varphi^0 = b$), with $b = \log(32)$. We refer to them as the “atypical” and “typical” ancestor scenarios, in the same order. In general, the profiles of the empirical statistics resemble the model statistics, with fluctuations that depend on the random terms η^v. In the case of an atypical ancestor, the mean displays transient dynamics that converge to b at exponential rate a, as per Eq. (3). For small enough a, convergence is attained within the experimental period, thus mean data suffice in principle for estimation of both a and b. For large a, instead, the steady-state value b is not apparent. In the case of a typical ancestor, the model mean provides the value of b, but it does not convey any information about a. In contrast, due to the small number of individuals in the early generations, the empirical mean may depart from b (to an extent dependent on ω^2), which is recovered within the experimental period for small a only. This happens because the observed individuals are related through their lineage tree, whence the empirical statistics are correlated across generations. If this is not taken into account, interpretation of mean data is deceptive and may lead to estimation bias. Instead, exploiting correlations may enable reconstruction of a even if the model mean dynamics are insensitive to it. Concerning variance profiles, given the single ancestor, transient dynamics are present in both scenarios. By Eq.(3), model variance grows from zero to $\omega^2/(1-a^2)$ at rate a^2. Thus, provided a relation between model and empirical variance to be established, variance dynamics always convey information about a and ω.

In conclusion, we saw that such simple model of trait evolution over a tree of dividing cells gives rise to dynamics of generation snapshot statistics that are more complex to interpret, but also potentially more informative, than population snapshot data collected on independent individuals [7]. How to appropriately process this data for model inference is the object of next section.

III. RECONSTRUCTION OF TRAIT EVOLUTION DYNAMICS FROM DIRECT MEASUREMENTS

We now consider inference of parameters $\theta = (A, \mathbf{b}, \Omega, \boldsymbol{\varphi}_0)$ from snapshot statistics $\tilde{\boldsymbol{\mu}}(n)$, $\tilde{\Sigma}(n)$, over N_g+1 generations, *i.e.* $n = 0, 1, \dots, N_g$. We will develop an exact maximum likelihood approach for inference from mean data, and an extension for the joint use of mean and variance data, and demonstrate their performance on simulated data. For all n, we assume $R(n)$ to be known (or estimated in a preliminary step, Sec. IV), and invertible.

A. Estimation from empirical means only

Let $\mathbf{y} = (\tilde{\boldsymbol{\mu}}(0)^{\mathbf{T}}, \tilde{\boldsymbol{\mu}}(1)^{\mathbf{T}}, \cdots, \tilde{\boldsymbol{\mu}}(N_g)^{\mathbf{T}})^{\mathbf{T}} \in \mathbb{R}^{m\times(N_g+1)}$. This is a Gaussian random vector whose mean $\bar{\mathbf{y}}_\theta = \mathbb{E}_\theta(\mathbf{y})$ and covariance matrix $\Gamma_\theta = \mathrm{Var}_\theta(\mathbf{y})$ have structure

$$\begin{pmatrix} \bar{\mathbf{y}}_\theta(0) \\ \vdots \\ \bar{\mathbf{y}}_\theta(N_g) \end{pmatrix}, \quad \begin{pmatrix} \Gamma_\theta(0,0) & \cdots & \Gamma_\theta(N_g,0)^{\mathbf{T}} \\ \vdots & \ddots & \vdots \\ \Gamma_\theta(N_g,0) & \cdots & \Gamma_\theta(N_g,N_g) \end{pmatrix}, \tag{6}$$

with $\bar{\mathbf{y}}_\theta(i) = \mathbb{E}_\theta(\tilde{\boldsymbol{\mu}}(i))$, $\Gamma_\theta(i,j) = \mathrm{Cov}_\theta(\tilde{\boldsymbol{\mu}}(i), \tilde{\boldsymbol{\mu}}(j))$, $\forall i, j$.

Proposition 1: For $0 \le j \le i \le N_g$, it holds that

$$\bar{\mathbf{y}}_\theta(i) = A^i\boldsymbol{\varphi}^0 + (I - A^i)\mathbf{b}\,,$$
$$\Gamma_\theta(i,j) = \frac{1}{2^{i\wedge j}}\left(A^{|i-j|}\sum_{k=0}^{i\wedge j-1} 2^k A^k \Omega A^{k^{\mathbf{T}}} + \delta_{i,j}R(i)\right),$$

with $\delta_{i,j}$ the Kronecker delta and $i \wedge j = \min\{i,j\}$.

The expression of $\bar{\mathbf{y}}_\theta(i)$ is equal to (3). The off-diagonal blocks of Γ_θ are nonzero as a result of correlation across generations. Factor $A^{|i-j|}$ in $\Gamma_\theta(i,j)$ represents the correlation decay of empirical means, the smaller the A, the faster the possible fluctuations along generations. Given $\mathbf{y}$, for a suitable parameter space Θ, we define $\hat{\theta}_{\mathbf{y}} \in \Theta$ as the maximum likelihood estimator of θ. Equivalently,

$$\hat{\theta}_{\mathbf{y}} = \arg\min_{\theta\in\Theta} \ell(\theta|\mathbf{y}) \tag{7}$$

where, up to an additive constant independent of θ, $\ell(\theta|\mathbf{y})$ is the negative log-likelihood function, given by

$$\ell(\theta|\mathbf{y}) = \frac{1}{2}\left(\log(|\Gamma_\theta|) + (\mathbf{y} - \bar{\mathbf{y}}_\theta)^{\mathbf{T}}\Gamma_\theta^{-1}(\mathbf{y} - \bar{\mathbf{y}}_\theta)\right)\,. \tag{8}$$

By Prop. 1, this function can be evaluated efficiently for any $\theta \in \Theta$, and the solution of (7) can be sought by numerical optimization. In practice, in view of (8), estimates $\hat{\theta}_{\mathbf{y}}$ result from the matching of mean dynamics (differences between data $\mathbf{y}$ and predictions $\bar{\mathbf{y}}_\theta$ must be small) and of correlations (differences $\mathbf{y} - \bar{\mathbf{y}}_\theta$ must agree with the structure of matrix Γ_θ). We thus fully profit from the information carried by empirical means and illustrated in the previous section.

B. Estimation from empirical means and variances

In principle, by extending the approach of the previous section, one could tackle inference from empirical mean and variance data by maximization of their joint likelihood. Unfortunately, deriving the joint likelihood for our tree-structured model is a formidable task. We therefore introduce a fitting cost function combining $\ell(\theta|\mathbf{y})$ with a suitable fitting term for the empirical variances over generations. For $n = 1, \dots, N_g$ let $\mathbf{v}(n)$ be a column vector defined as $\mathbf{v}(n) = \mathscr{S}\big(\tilde{\Sigma}(n)\big)$, where $\mathscr{S}$ is a linear operator extracting suitable elements of $\tilde{\Sigma}(n)$, and let $\bar{\mathbf{v}}_\theta(n) = \mathbb{E}_\theta(\mathbf{v}(n))$. For $\mathbf{v}$ the (vector) collection of the $\mathbf{v}(n)$, we define the cost

$$c(\theta|\mathbf{y},\mathbf{v}) = \ell(\theta|\mathbf{y}) + \frac{1}{2}\sum_{n=1}^{N_g}\tilde{\mathbf{v}}_\theta(n)^{\mathbf{T}}W_\theta^{-1}(n)\tilde{\mathbf{v}}_\theta(n)\,, \tag{9}$$

with $\tilde{\mathbf{v}}_\theta(n) = \mathbf{v}(n) - \bar{\mathbf{v}}_\theta(n)$, and the estimator of θ from empirical mean and variance data as

$$\hat{\theta}_{\mathbf{y},\mathbf{v}} = \arg\min_{\theta\in\Theta} c(\theta|\mathbf{y},\mathbf{v}). \tag{10}$$

In analogy with the first term $\ell(\theta|\mathbf{y})$, the second term amounts to squared residuals between empirical (variance) statistics and their expected value, weighted by suitable matrices $W_\theta(n)$ that must ensure an appropriate tradeoff between the fitting cost for means (first term) and variances (second term). We will discuss the choice of the $W_\theta(n)$ soon.

Proposition 2: For $n = 0, \dots, N_g$ it holds that

$$\mathbb{E}_\theta\left(\tilde{\Sigma}(n)\right) = \Sigma_\theta(n) + R(n) + \sum_{i=0}^{n-1}\frac{1-2^i}{2^n-1}A^i\Omega A^{i^{\mathbf{T}}} \tag{11}$$

and $\bar{\mathbf{v}}_\theta(n) = \mathscr{S}\left(\mathbb{E}_\theta\left(\tilde{\Sigma}(n)\right)\right)$, with $\Sigma_\theta(n)$ given by (3).

Eq. (11) shows that $\tilde{\Sigma}(n)$ has expected value equal to $\Sigma(n)$ plus ($R(n)$ and) a nontrivial term that results from the correlation of individuals across generations. Failing to account for this term in a naive match between empirical and model variances would introduce bias. Instead, by virtue of Eq. (11), the second term in (9) duly quantifies deviations between empirical and model variance dynamics.

Inspired from $\ell(\theta|\mathbf{y})$, provided a suitable definition of $\mathscr{S}$ ensuring invertibility, one may define $W_\theta(n) = \mathrm{Cov}_\theta(\mathbf{v}(n))$. Calculating this covariance matrix is generally complex. We instead approximate it by its expression for $A = 0$

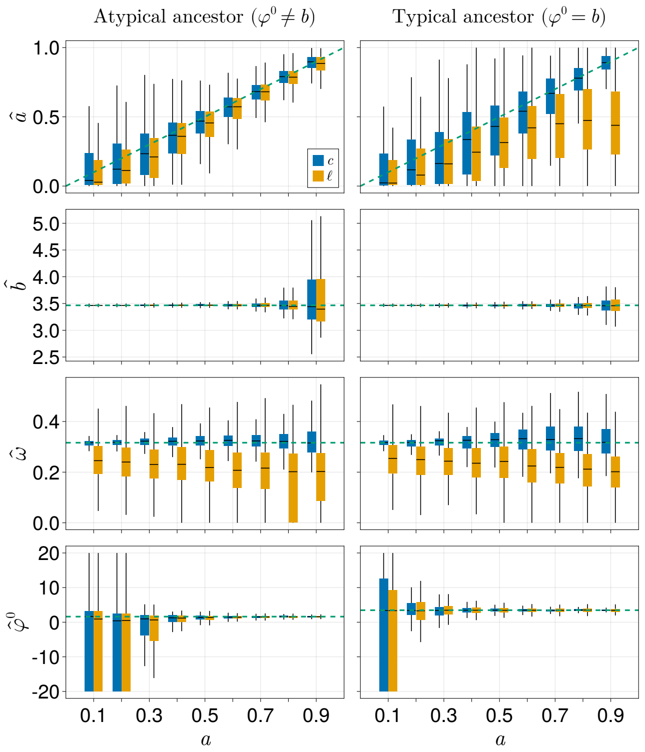

Fig. 2. Estimation performance for different simulated values of a from direct measurements of individual parameters in the atypical (left) and typical (right) ancestor scenarios, using means only (Eq. (7), yellow), and also using variances (Eq. (10), blue). True values: Dashed green line

(uncorrelated individuals within generations). In this case, for indices k and k' such that $\mathscr{S}$ maps element (i,j) (resp (i',j')) of its input matrix into element k (resp. k') of its output vector, the element in position (k,k') of matrix $W_\theta(n)$ is given by $(|S_n|-1)^{-1}(\varsigma_{i,i'}\varsigma_{j,j'}+\varsigma_{i,j'}\varsigma_{j,i'})$ (see [6, Thm. 3.3.15]), where $\varsigma_{i,j}$ is the element of position (i,j) of matrix $\Sigma_\theta(n)+R(n)$. Using this expression and Eq. (11), cost (9) can be evaluated efficiently for any θ, and the solution of problem (10) can be sought by numerical optimization. The effectiveness of our approach is demonstrated next.

C. Numerical performance assessment

We now evaluate performance of the estimation methods developed above in the two scenarios of atypical and typical ancestor. Fig. 2 reports the results from Monte-Carlo analysis based on simulated datasets. One dataset consists in one random realization of Eq. (1) over a complete binary tree with 10 generations ($N_g = 9$). Noisy measurements of the simulated parameters are generated in accordance with Eq. (2). Simulation parameters are $b = \log(32)$, $\omega^2 = 0.1$ and $R(n)$ equal to $r^2 = 0.1$ for all n. For every value $a = 0.1, 0.2, \ldots, 0.9$, we generated 500 such datasets for the typical ancestor scenario $\boldsymbol{\varphi}_0 = b$ and 500 datasets for the atypical ancestor scenario $\boldsymbol{\varphi}_0 = \log(5)$. For each simulated case, boxplots summarize the estimation results obtained from the application of the methods of Sec. III-A (usage of empirical means only) and of Sec. III-B (usage of empirical means and variances, with $\mathscr{S}$ set to the identity map) to each of the 500 datasets. In our implementation in Julia, every estimation run takes 1–2 seconds on a modern laptop.

Results reflect the expectations of Sec. II. In the atypical ancestor scenario, using empirical means only yields estimates of a that appear unbiased, and more accurate for larger values of a (transient mean dynamics not exhausted within the very first generations, where empirical means are noisier). The same is true for the estimates of φ_0, which are essentially undefined for small a. Estimates of b instead worsen with increasing a because of the limited number of observed generations. Estimates of ω, although biased, are of the right order of magnitude. This is an interesting consequence of the fact that ω enters problem (7) via Γ_θ. The additional use of empirical variances leads to analogous results, the only exception being the improved estimates of ω. This is explained by the role of ω in Eq. (11). In the typical ancestor scenario, the interest of exploiting empirical variances to guarantee better estimates of a becomes apparent. Nonetheless, estimates of a from means only overall follow the values of a as they increase. This is quite remarkable and entirely due to the covariance matrix Γ_θ entering problem (7). For both methods, performance in the estimation of b, φ_0, and ω, is qualitatively comparable to the atypical ancestor scenario.

In summary, two things were shown. First, accounting for correlations among empirical means enables estimation of the model parameters from means only, although estimates of ω and a are biased in some cases. This is conceptually interesting and it may have practical relevance for applications where only empirical means are available. Second, further exploiting empirical variances enables estimation of all model parameters in all conditions. This finding qualifies optimization (10) as an effective parameter estimation method, and leads us to focus on this method in the sequel.

IV. Reconstruction from indirect parameter measurements: Gene expression case study

So far, we have developed methods to reconstruct parameters $\theta = (A, \mathbf{b}, \Omega, \boldsymbol{\varphi}_0)$ based on noisy measurements $\tilde{\boldsymbol{\varphi}}^v$ of the individual-cell parameters $\boldsymbol{\varphi}^v$ with known variances $R(n)$. We now discuss their application to problems where measurements of individual-cell parameters are not directly available, and variances $R(n)$ are a priori unknown.

Motivated by [11], we do so for the case study of gene expression dynamics. In [11], expression of an osmosensitive gene in response to osmotic shocks is monitored over time in individual yeast cells by the use of a fluorescent reporter protein and videomicroscopy. Data is used to investigate variability of kinetic rate parameters of gene expression across cells. In [11], correlation of the estimated parameters across mother and daughter cells is only evaluated *a posteriori* based on the few cells for which parental relationships are available. In our more recent work [12], the mother-daughter cell correlation model (1) is identified directly, again limited to the cells with known parental relationships. The methods presented in Sec. III enable identification of the mother-daughter correlation model without knowledge of parental relationships. To achieve this, however, individual-cell parameter estimates $\tilde{\boldsymbol{\varphi}}^{v_n}$ and variances $R(n)$ must be obtained for every generation n in a preliminary step.

In Sec. IV-A we elaborate on this preliminary step. We propose robust methods to obtain individual-cell parameters $\tilde{\boldsymbol{\varphi}}^{v}$ and their variances $R(n)$ from gene expression time profiles. In conjunction with the method of Sec. III-B, this results in a pipeline to estimate parameters θ from gene expression time profiles in absence of lineage information. In Sec. IV-B, we show performance on simulations of the experiments of [11]. While focused on the gene expression case study, the methods proposed can be readily generalized.

A. Inference from single-cell gene expression profiles

The procedure to obtain individual-cell parameters $\tilde{\boldsymbol{\varphi}}^{v}$, with $v \in S_n$, and $R(n)$ from gene expression time profiles applies separately to every generation n. We assume that the generation a cells belongs to is known (see comments in Sec. V). For a fixed n, let $t_0, \ldots, t_{N_v-1}$ be increasing measurement times for cell $v \in S_n$. Let $\tilde{p}_1^v, \ldots, \tilde{p}_{N_v}^v$ be measurements of the cellular concentration of a reporter protein obeying $\tilde{p}_\ell^v = p_\ell^v + \varepsilon_\ell$, with p_ℓ^v the true concentration at time t_ℓ and $\varepsilon_\ell \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ measurement noise uncorrelated across time and cells. Variance σ_ε^2 is considered unknown. We assume that $p_\ell^v = p(t_\ell|\boldsymbol{\varphi}^v)$, where $p(t|\boldsymbol{\varphi}^v)$ is the solution at time t of the simple gene expression model

$$\frac{\mathrm{d}}{\mathrm{d}t} p(t) = -\gamma p(t) + \kappa^v u(t), \quad p^v(t_0^v) = p_0^v, \qquad (12)$$

with $\boldsymbol{\varphi}^v = (\kappa^v, p_0^v)$. Concentration decay rate γ results from growth dilution and protein degradation, κ^v is the protein synthesis rate upon gene expression activation, and $u(t)$ represents activation ($u = 1$) and deactivation ($u = 0$) at time t in response to known exogenous stimuli, assumed identical across cells [11]. For stable reporter proteins, this simple model is a viable approximation of transcription-translation dynamics, and γ is determined by growth rate uniformly across cells (see [4] and references therein). Among the entries of $\boldsymbol{\varphi}^v$, we will focus on the evolution over generations of κ^v. That is, model (1) will be restricted to the scalar $\varphi^v = \kappa^v$. No transmission model is postulated for protein concentrations. Yet we let p_0^v differ across cells, *i.e.* it is part of the unknown individual-cell parameters $\boldsymbol{\varphi}^v$.

Estimation of $\boldsymbol{\varphi}^v$ from data $\mathcal{D}^v = \{(t_\ell^v, \tilde{p}_\ell^v),\ \ell = 0, \ldots, N_v\}$ could be performed by least-squares fitting separately for every cell v, however, performance may be limited for sparse data. We propose instead a Mixed-Effects (ME) approach [10]. According to the basic ME paradigm, the unknown parameters $\boldsymbol{\varphi}^v$ of all individuals of a given population (for us, S_n) are treated as random outcomes from a common distribution $\mathcal{F}(\Xi)$ with parameters Ξ. Given knowledge of the individual statistical response model $p(\cdot|\boldsymbol{\varphi}^v)$, and the model relating noisy measurements $\tilde{p}_\ell^v$ with $p(\cdot|\boldsymbol{\varphi}^v)$, data $\mathcal{D}^v$ from all individuals $v \in S_n$ are pooled together to calculate an estimate of Ξ, along with estimates $\hat{\boldsymbol{\varphi}}^v$ of individual parameters $\boldsymbol{\varphi}^v$, and an estimate $\hat{\sigma}_\varepsilon^2$ of σ_ε^2. Thanks to treating individuals as part of a same statistical population, ME inference is known to outperform individual inference especially for noisy, short individual time series [10]. We will specifically rely on the ME inference method known as SAEM [10], assuming that $\mathcal{F}(\Xi)$ is a log-normal distribution to complete the problem specification.

Suppose that estimates $\{\hat{\boldsymbol{\varphi}}^v : v \in S_n\}$ and $\hat{\sigma}_\varepsilon^2$ have been obtained by ME inference. To enable application of the method of Sec. III-B, we set $\tilde{\boldsymbol{\varphi}}^v = \hat{\boldsymbol{\varphi}}^v$ for all v, and deduce $R(n)$ from $\hat{\sigma}_\varepsilon^2$ as follows. For every individual v, the variance Υ^v of the parameter estimate $\hat{\boldsymbol{\varphi}}^v$ is approximated by $\Upsilon^v = \hat{\sigma}_\varepsilon^2 (G_v^{\mathbf{T}} G_v)^{-1}$, where the ℓth row of matrix G_v, defined by $[\partial p(t_\ell|\boldsymbol{\varphi}^v)/\partial \boldsymbol{\varphi}^v]_{\boldsymbol{\varphi}^v = \hat{\boldsymbol{\varphi}}^v}$, is the sensitivity of $p(t_\ell|\boldsymbol{\varphi}^v)$ to variations in $\boldsymbol{\varphi}^v$. G_v is easily calculated *e.g.* by the sensitivity equations [8]. Finally, $R(n)$ is defined as the mean of Υ^v across all $v \in S_n$. In sums, the whole pipeline goes as follows. For $n = 0, \ldots, N_g$: *(i)* Given data $\mathcal{D}^v$ for all $v \in S_n$, calculate estimates $\{\hat{\boldsymbol{\varphi}}^v : v \in S_n\}$ and $\hat{\sigma}_\varepsilon^2$ by SAEM; *(ii)* For every individual $v \in S_n$ run sensitivity equations to calculate G_v and $\Upsilon^v = \hat{\sigma}_\varepsilon^2 (G_v^{\mathbf{T}} G_v)^{-1}$; *(iii)* Calculate $R(n)$ as the mean of Υ^v across $v \in S_n$, and set $\tilde{\boldsymbol{\varphi}}^v = \hat{\boldsymbol{\varphi}}^v$ for all $v \in S_n$; *(iv)* Calculate empirical statistics (4)-(5). Finally, find $\hat{\theta}$ by solving problem (10). This last step is readily modified to focus on some entries of $\boldsymbol{\varphi}^v$ only, as it is the case in next section. Note that estimates of $\tilde{\mu}(n)$ and $\tilde{\Sigma}(n)$ are also inherently calculated by SAEM in the reconstruction of Ξ. Yet, their construction in SAEM may not fulfill Eq. (11).

B. Simulation results

We now show performance of the estimation pipeline by a numerical Monte-Carlo study. We consider the gene expression model (12), with parameter $\varphi^v = \kappa^v$ obeying the AR model (1) and γ fixed to 0.01. We consider cells v over $N_g + 1 = 10$ generations and assume that all cells of generation n are born at time $100n$ and divide at time $100(n+1)$ (minutes). For parameters b and ω fixed as in Sec. III-C, in the typical ancestor scenario ($\boldsymbol{\varphi}^0 = b$), we generated 50 gene expression datasets for each value $a = 0.1, \ldots, 0.9$. Every dataset is obtained in three steps: Simulation a complete binary tree of parameter values κ^v over $N_g + 1 = 10$ generations, as per Eq. (1); Simulation of the response (12) for every individual κ^v, with initial condition p_0^v fixed to the mother cell final concentration $p^{v^-}(t_0^v)$ (for the common ancestor the initial condition is set to zero); Simulation of the noisy measurements (12), with noise strength realistically set to $\sigma_\varepsilon = 20$. Fig. 3 illustrates the chosen input profile $u(t)$ and example simulated data over one branch of one population tree. For every value of a, we then run the estimation pipeline of Sec. IV-A on each of the 50 simulated datasets. Due to difficulties encountered with the SAEM implementations in Julia, we implemented the first part of the pipeline in Matlab [13] based on function `nlmefitsa`. One complete run of the pipeline for one population tree takes about 5 minutes.

Estimation statistics are shown in Fig. 4. Results are qualitatively similar to those of Sec. III-C, which were based on direct parameter measurements. Estimates remain unbiased and reasonably concentrated around the true values

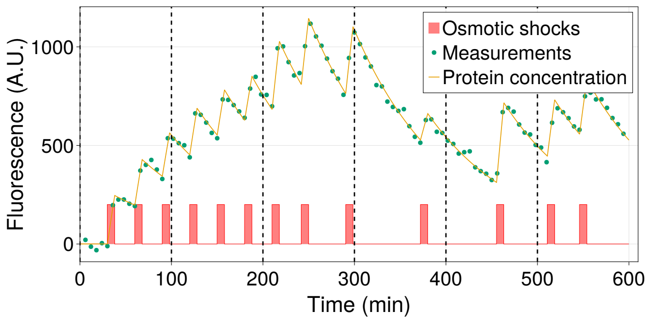

Fig. 3. Example simulation of gene expression dynamics along one branch of the population tree. Vertical dashed lines: Cell division times.

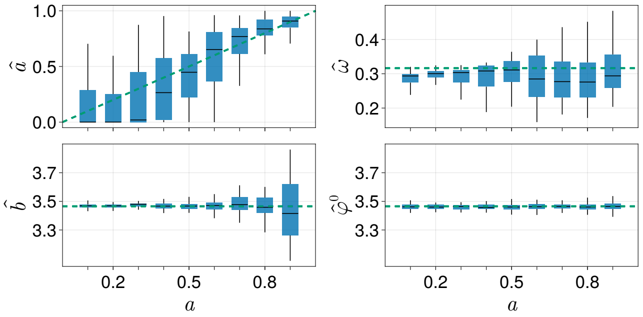

Fig. 4. Estimation performance from gene expression profiles, for different simulated values of a. True values: Dashed green lines

in all cases. Quantitative comparison with the results in Fig. 2 is not appropriate, since the error variances $R(n)$ that enter the results of Fig. 4 are estimated from the simulated data and generation-dependent. The extent to which estimation of matrices $R(n)$ contributes to overall estimation uncertainty will be assessed in a future study.

V. CONCLUSIONS AND FUTURE WORK

We have addressed estimation of an AR model of individual-cell trait evolution on a population tree from trait statistics within generations. We developed general methods for arbitrary traits measured directly, and a specific yet generalizable extension for kinetic rate parameters measured indirectly in terms of single-cell gene expression profiles. We showed that the tree-structured correlation of the parameters plays a crucial role in the definition of appropriate estimators, and demonstrated performance in simulations directly related with real experiments from [11]. Our contributions are of direct interest to single-cell microscopy data, and they provide a step into the broader problem of treating snapshot measurements from tree-structured populations. While our methods have been showcased relative to evolution of scalar (or uncorrelated) parameters, they are applicable to vectors of correlated parameters. Application to the data in [11] and comparison of results with those assuming lineage information of [12] is in progress. In so doing we are addressing the assumption that the generation that cells belong to is known. Unreported efforts show that the experimental data can be reconciled with this assumption, with encouraging results. Beyond the biological case study, our contribution is of potential interest to any application with tree-structured data. Future directions of research include theoretical assessment of identifiability and estimation performance, extension of the methods to partially observed trees, and developments toward snapshot data not aligned with generations.

VI. ACKNOWLEDGMENTS

The authors thank Jakob Ruess for useful discussions.

REFERENCES

[1] C. Aditya, F. Bertaux, G. Batt, and J. Ruess. Using single-cell models to predict the functionality of synthetic circuits at the population scale. *PNAS*, 119(11):e2114438119, 2022.

[2] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A Fresh Approach to Numerical Computing. *SIAM Review*, 59(1):65–98, 2017.

[3] E. Y. Bijman, H.-M. Kaltenbach, and J. Stelling. Experimental analysis and modeling of single-cell time-course data. *Curr. Opin. Syst. Biol.*, 28:100359, 2021.

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

[5] L. Duso and C. Zechner. Stochastic reaction networks in dynamic compartment populations. *PNAS*, 117(37):22674–22683, 2020.

[6] A. K. Gupta and D. K. Nagar. *Matrix Variate Distributions*. Chapman and Hall/CRC, 1999.

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

[8] H.K. Khalil. *Nonlinear Systems*. Pearson Education. Prentice Hall, 2002.

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

[10] Marc Lavielle. *Mixed Effects Models for the Population Approach Models, Tasks, Methods and Tools*. Chapman and Hall/CRC, New York, 2014.

[11] A. Llamosi, A. M. Gonzalez-Vargas, C. Versari, E. Cinquemani, G. Ferrari-Trecate, P. Hersen, and G. Batt. What Population Reveals about Individual Cell Identity: Single-Cell Parameter Estimation of Models of Gene Expression in Yeast. *PLOS Comput. Biol.*, 12(2):e1004706, 2016.

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

[13] MATLAB. *version 9.13 (R2022b)*. The MathWorks Inc., Natick, Massachusetts, 2010.

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

[15] A. Raj and A. van Oudenaarden. Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences. *Cell*, 135(2):216–226, 2008.

[16] S. Taheri-Araghi, S. Bradde, J. T. Sauls, N. S. Hill, P. A. Levin, J. Paulsson, M. Vergassola, and S. Jun. Cell-Size Control and Homeostasis in Bacteria. *Curr. Biol.*, 25(3):385–391, 2015.

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

[18] N. Totis, C. Nieto, A. Kuper, C. Vargas-Garcia, A. Singh, and S. Waldherr. A Population-Based Approach to Study the Effects of Growth and Division Rates on the Dynamics of Cell Size Statistics. *IEEE Control Syst. Lett.*, 5(2):725–730, 2021.

[19] S. Waldherr. Estimation methods for heterogeneous cell population models in systems biology. *J. R. Soc. Interface*, 15(147):20180530, 2018.

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

[21] C. Zechner, M. Unger, S. Pelet, M. Peter, and H. Koeppl. Scalable inference of heterogeneous reaction kinetics from pooled single-cell recordings. *Nat. Methods*, 11:197–202, 2014.