2013 European Control Conference (ECC)
July 17-19, 2013, Zürich, Switzerland.

Identification of biological models from single-cell data: a comparison between mixed-effects and moment-based inference

Andres M. Gonzalez[†], Jannis Uhlendorf, Joé Schaul, Eugenio Cinquemani, Gregory Batt and Giancarlo Ferrari-Trecate [*]

***Abstract*— Experimental techniques in biology such as microfluidic devices and time-lapse microscopy allow tracking of the gene expression in single cells over time. So far, few attempts have been made to fully exploit these data for modeling the dynamics of biological networks in cell populations. In this paper we compare two modeling approaches capable to describe cell-to-cell variability: Mixed-Effects (ME) models and the Chemical Master Equation (CME). We discuss how network parameters can be identified from experimental data and use real data of the HOG pathway in yeast to assess model quality. For CME we rely on the identification approach proposed by Zechner *et al.* (PNAS, 2012), based on moments of the probability distribution involved in the CME. ME and moment-based (MB) inference will be also contrasted in terms of general features and possible uses in biology.**

I. Introduction

It is now well recognized that the functioning of biological systems at the molecular and cellular levels is noisy [1]. Within a genetically-identical population, cells behave in a heterogeneous and stochastic manner. The biological role of noise, whether beneficial or detrimental, is still a major open question in systems biology [2], therefore, models that are able to capture various aspects of cell population's heterogeneous behaviors would shed light on the matter.

In short, one distinguishes two major origins of noise: intrinsic noise, coming from stochasticity of the observed process itself, and extrinsic noise, coming from differences between cells or cell's environment [1]. Arguably, the modeling framework of choice for intrinsic variability is the use of CME. Unfortunately, the identification of such models is quite difficult in practice, although approximations of the CME [3] have been recently proposed and used to fit probability distributions predicted by the model to empirical distributions from cell populations. Alternatives for the identification of CME model parameters, using moment-closure methods [4], have also proved to be biologically relevant and computationally much more tractable [5]; however, in this framework, only some statistics of the population's distribution are used. These modeling approaches are, therefore, well adapted to flow cytometry data, which provides information on population distributions, but are unable to exploit the rich information on single-cell temporal evolution that time-lapse movie data provide. For this kind of data, we propose to use ME models. In this framework, a population of individuals is described by an ODE model supplemented by parameter distributions. Importantly, in contrast to CME and MB approaches, ME models naturally capture extrinsic variability.

As a statistical analysis framework, ME has been previously used in fields such as medicine, ecology, manufacturing, psychology and, especially, it has been for decades the modeling paradigm in the field of pharmacokinetics [6], [7]. However, to the best of our knowledge, the identification of ME models of cellular processes based on single cell fluorescence microscopy data has not yet been studied.

In this paper, we compare ME models with CME models identified through an MB approach very similar to the one presented in [5], using fluorescence microscopy data that shows the response of yeast cells to repeated osmotic shocks in control experiments [8]. Our results show that methods for identification of ME models based on real video-microscopy data are computationally tractable, and that in comparison to state-of-the-art MB identification methods perform in a comparable way. We therefore conclude that based on our results, one cannot unambiguously decide whether the major source of variability in the observed cellular process is intrinsic or extrinsic noise. Lastly, we note that the performance of ME inference methods could be further improved by taking cell lineage, cell physiology, or other covariate information into account.

The paper is structured as follows: In section II we review mixed-effect models, CME models, and moment-closure methods. The biological system under study and the corresponding ME and MB models used in the inference process are described in section III. The quality of the identified models is analized in section IV. Finally, section V provides a critical comparison of ME and MB inference procedures. Through the paper we will use bold letters to denote random variables.

II. Model inference approaches

This section is a short review of the approaches that we will compare. In particular, we will detail the underlying assumptions, data required and output provided.

[*] The research leading to these results has received funding from the European Union Seventh Framework Programme [FP7/2007-2013] under grant agreement n 257462 HYCON2 Network of Excellence, the INRIA/INSERM project Colage, and the French National Research Agency research grant Iceberg ANR-IABI-3096

A.M. Gonzalez and G. Ferrari-Trecate are with Dipartimento di Ingegneria Industriale e dell'Informazione, Università degli Studi di Pavia, Pavia, Italy

J. Uhlendorf, J, Schaul and G. Batt are with INRIA Paris-Rocquencourt, France

E. Cinquemani is with INRIA - Grenoble Rhône-Alpes, France

[†]Corresponding author

andres.gonzalez01@universitadipavia.it

A. Mixed-Effects model inference

ME modeling is an approach to the analysis of population data where samples are clustered into subgroups of the population, called "individuals", in which the observations are mutually dependent (e.g. time-series of single cell's fluorescence) and the behavior of the total population is inferred from these individuals. The general form of an ME model is [7], [6]

$$\beta_i = d(\alpha_i, \beta, \boldsymbol{b_i}), \quad i = 1, \ldots, N \tag{1}$$

$$\boldsymbol{y_{ij}} = f(Z_{ij}, \beta_i) + \boldsymbol{e_{ij}}, \quad j = 1, \ldots, J \tag{2}$$

Equation (2) represents the model of the ith individual, where $\boldsymbol{y_{ij}} \in \mathbb{R}$ denotes the jth measure of individual i, defined by f, a function that depends on a set of regressors $Z_{ij} \in \mathbb{R}^N$, and a vector of individual-specific parameters, $\beta_i \in \mathbb{R}^p$. Measurement noise is represented by the independent random variable $\boldsymbol{e_{ij}}$. Further considerations can be made at this level, for example, when dealing with time series of data. In this case, the vector of regressors can be expressed as $Z_{ij}=(t_{ij}, u_{ij})$, where $t_{ij} \in \mathbb{R}$ and $u_{ij} \in \mathbb{R}$ are the time and input of the ith cell at jth instant. The parameters β_i are extracted from (1), the population model, where d is a function of α_i, a vector of "covariates" (known individual factors), $\beta \in \mathbb{R}^p$, a vector of "fixed effects", and $\boldsymbol{b_i}$, a vector of "random effects", which is characterized by a covariance matrix $C \in \mathbb{R}^{p \times p}$. The population function d in (1) and statistical assumptions on $\boldsymbol{b_i}$ induce a probability distribution on the parameters. The classical methods to estimate d, as well as some statistics of $\boldsymbol{b_i}$, are based on likelihood maximization [6].

B. Moment-Based inference

A standard approach for a detailed description of (bio)chemical reaction kinetics relies on probabilistic modelling of the reactions among discrete pools of molecules. Consider a reaction network involving n species taking part in R reactions. Denote with $\mathbf{x}_s(t)$, where $s = 1, \ldots, n$, the number of molecules of the sth species at time t. For $r = 1, \ldots, R$, let $\nu_{r,s} \in \mathbb{Z}$ be the change in the number of molecules of the sth species upon occurrence of the rth reaction (i.e., signed stoichiometric coefficients), and let $a_r(x) \cdot \delta t$, with $x = (x_1, \ldots, x_n) \in \mathbb{N}^n$ and $a_r(x) \geq 0$, be the probability that the reaction occurs in the infinitesimal time δt given x molecules of the various species. Then $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ can be seen as the state of a continuous-time, discrete-jump Markov process taking values in $\mathbb{N}^n$ [9]. The probability $p(x,t)$ that $\mathbf{x}$ is equal to x at time t evolves according to the so-called Chemical Master Equation (CME)

$$\frac{d}{dt}p(x,t) = \sum_{r=1}^{R} a_r(x - \nu_r)p(x - \nu_r, t) - a_r(x)p(x,t), \tag{3}$$

where $\nu_r = (\nu_{r,1}, \ldots, \nu_{r,n})$. The reaction propensities $a_r(x)$ are typically fixed by the the laws of mass-action [10] up to a vector of kinetic parameters. As an alternative characterization of the probability law $p(x,t)$, one may consider the collection of all moments $\mu_\ell(t) = \mathbb{E}[\mathbf{x}^\ell]$, where $\ell = (\ell_1, \ldots, \ell_n) \in \mathbb{N}^n$ is a multi-index, $\mathbf{x}^\ell = \mathbf{x}_1^{\ell_1} \cdot \ldots \cdot \mathbf{x}_n^{\ell_n}$, and $|\ell| = \ell_1 + \ldots + \ell_n$ is the order of μ_ℓ.

Excluding certain special cases, neither the solution of (3) nor the related moments of $\mathbf{x}$ can be computed in practice, therefore several approximation techniques have been developed; one of such techniques is Moment Closure (MC). For propensities $a_r(x)$ polynomial in x (as in mass-action kinetics) and some $L \in \mathbb{N}_{>0}$, let μ denote a vector containing all moments up to order L. For matrices A and B fixed by (3) and a suitable so-called moment closure function $\phi(\cdot)$, an approximation $\tilde{\mu}(t)$ of $\mu(t)$ can be obtained through the system [4]

$$\frac{d}{dt}\tilde{\mu}(t) = A\tilde{\mu}(t) + B\phi\big(\tilde{\mu}(t)\big). \tag{4}$$

Models of this form have been used in [5], [9] to describe the dynamics of regulatory networks within single cells of a cell population. Let us denote with θ the vector containing the kinetic parameters of the propensity functions and the statistics of the noise (see (6)). Note that θ (as well as the system statistics at an initial time) may vary across cells. Assume first that θ is unknown but fixed across the population. Moment-Based Inference (see e.g. [5]) assumes that parameters θ are estimated by fitting the solutions of equations like (4) to the corresponding empirical moments of experimental data. In particular, let $c^T\mathbf{x}$, with $c \in \mathbb{R}^n$, be an observed scalar output of a network (e.g. the expression of one gene), and assume that, for every cell in a population of N cells, measurements $\mathbf{y}$ follow the model

$$\mathbf{y} = c^T\mathbf{x} + \mathbf{e}, \tag{5}$$

$$\mathbf{e} = \big(e_a + e_b c^T\mathbf{x}\big)\boldsymbol{\eta}, \tag{6}$$

with $\boldsymbol{\eta} \sim \mathcal{N}(0,1)$. Mean m_y and second-order uncentered moment M_y of y relate to the moments $m_x \simeq \mu_1$ and $M_x \simeq \mu_2$ of x by way of the equations

$$m_y = c^T m_x, \tag{7}$$

$$M_y = (1 + e_b^2)\mathrm{var}(c^T x) + (e_a + e_b c^T m_x)^2 + (c^T m_x)^2 \tag{8}$$

At any given measurement time, empirical versions $\hat{m}_y$ and $\hat{M}_y$ of m_y and M_y can be determined from cell population histograms. Then, an estimate of the model parameters θ may be defined by the solution of the optimization problem

$$\min_\theta \quad \sum_j D\big(\hat{m}_{y,j}, \hat{M}_{y,j} || m^\theta_{y,j}, M^\theta_{y,j}\big), \tag{9}$$

where a superscript symbol θ denotes dependence on the parameters, $\hat{m}_{y,j}, \hat{M}_{y,j}, m_{y,j}, M_{y,j}$ denote the corresponding moments at time t_j, and D denotes a suitable distance. In practice, m^θ_y and M^θ_y are computed based on the solutions of a moment-closure equation (4) with A, B (and possibly ϕ) depending on θ.

Similar to what is done for reaction rate (ODE) models (see e.g. [11], and Section III-B.2), the case where θ takes different values depending on the subject (extrinsic noise) could be included in this framework by assigning priors to θ

with hyperparameters common to the whole population [5]. In this case, for model inference, the process statistics μ_ℓ must also account for this additional source of variability, and the hyperparameters of the prior, rather than the individual random values of θ, have to be estimated [11]. In [5], an alternative approach was proposed in particular for the case study of this paper, where certain extrinsic noise quantities where treated as state invariants of a CME-like model. This allowed to confine the uncertainty of these quantities into an appropriately chosen initial distribution, and to compute the evolution of the augmented system moments in a way at all similar to (4).

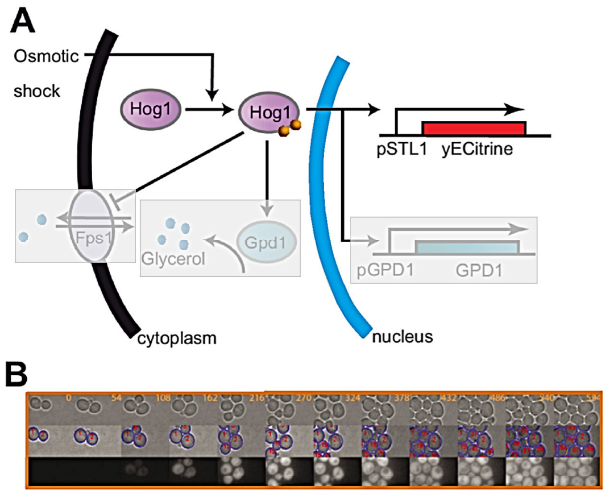

Fig. 1. Hyperosmotic gene expression in yeast. (A) Hyperosmotic stress triggers phosphorylation and nuclear import of the protein Hog1,which thereupon activates osmo-stress responsive genes. In addition Hog1 stimulates enzymes involved in the glycerol production pathway, while closure of the membrane glycerol transporter Fps1 prevents glycerol from leaking out. Increasing the intracellular glycerol concentration is the main adaptation mechanism to hyperosmotic stress. Adaptation is prevented by our experimental setup, thus Fps1 and GPD1 mechanisms (depicted in light gray) are not considered in our model (B). Information gathered by fluorescence microscopy. Cells are grown in a microfluidic device which can select between normal and high osmolarity media. A microscope takes fluorescent images of the cells, which are segmented and tracked in real-time.

III. Case Study: Gene Expression in Yeast Cell Populations

A. Description of the system

In the budding yeast *S. cerevisiae*, an increase of the environmental osmolarity activates the high osmolarity glycerol (HOG) signal transduction pathway, a stress response pathway that coordinates the adaptation response to an osmotic shock. Adaptation is achieved by increasing the cellular glycerol level via various mechanisms, one of which is the upregulation of genes involved in glycerol production [12]. Here, we used the promoter of an osmoreponsive gene (*STL1*) to drive the expression of a yECitrine fluorescent protein (see Figure 1A) to monitor the gene expression response of the cells to repeated osmotic stresses. Recently, some of the authors presented a feedback control platform, utilizing the HOG pathway to control gene expression [8] and the data used here for model parametrization and evaluation stems from this work. Figure 1B illustrates the visual information gathered with this method.

B. Models under comparison

1) Underlying reaction network: The system is described by the reaction network of HOG1-induced gene expression (10) as proposed in [5] where the (delayed) gene activation rate, $u(t-\tau)$, caused by an osmotic shock, represents the system's input, and the amount of fluorescent protein yECitrine is the output.

$$\begin{aligned} pSTL1^{off} &\underset{c_2}{\overset{c_1 u(t-\tau)}{\rightleftharpoons}} pSTL1^{on} \\ pSTL1^{on} + CR &\underset{c_4}{\overset{c_3}{\rightleftharpoons}} CR{\cdot}pSTL1^{on} \\ CR{\cdot}pSTL1^{on} &\xrightarrow{c_5} CR{\cdot}pSTL1^{on} + mRNA \\ mRNA &\xrightarrow{c_6} mRNA + yECitrine \\ yECitrine &\xrightarrow{c_7} \emptyset \\ mRNA &\xrightarrow{c_8} \emptyset \end{aligned} \tag{10}$$

In (10), $pSTL1^{on}$ and $pSTL1^{off}$ represent the active and inactive state of pSTL1 promoter (with rates of activation/deactivation c_1 and c_2). CR is the concentration of chromatin remodeling complex, which binds/unbinds to pSTL1 with rates c_3 and c_4 to produce the $CR{\cdot}pSTL1^{on}$ complex. The complex starts the production of mRNA at a rate c_5, which in turn produces yECitrine, with a synthesis rate c_6 that depends on the number of ribosomes and a kinetic parameter. Finally, yECitrine and mRNA are degraded with rates c_7 and c_8.

2) Model for ME inference: Next, we will characterize the functions f in (2) and d in (1). In our case, individuals correspond to single cells whose fluorescence is tracked over time. By the laws of mass-action, the (average) system can be described (as illustrated in [13]) by the set of reaction rate equations (11):

$$\begin{aligned} \dot{x_1} &= c_2 x_2 - c_1 u(t-\tau) x_1 \\ \dot{x_2} &= c_1 u(t-\tau) x_1 - c_2 x_2 + c_4 x_4 - c_3 x_2 x_3 \\ \dot{x_3} &= c_4 x_4 - c_3 x_2 x_3 \\ \dot{x_4} &= c_3 x_2 x_3 - c_4 x_4 \\ \dot{x_5} &= c_5 x_4 - c_8 x_5 \\ \dot{x_6} &= c_6 x_5 - c_7 x_6 \end{aligned} \tag{11}$$

where x_1, x_2, x_3, x_4, x_5, and x_6 represent, respectively, the proportion of the promoters that are in the *off* state ($pSTL1^{off}$), in the *on* state ($pSTL1^{on}$), or bound to chromatin remodeling factors ($CR{\cdot}pSTL1^{on}$), and the concentrations of chromatin remodeling factors (CR), $mRNA$ and fluorescent proteins yECitrine. The parameters to be identified are: reaction rates $c=\{c_1,c_2,c_3,c_4,c_5,c_6,c_7,c_8\}$, the initial concentration of CR ($x_{3(0)}$) and the time delay $\tau>0$. The procedure for calculating the input u is developed in [14]. For these ten parameters we want to estimate a mean value β and a covariance matrix C. In order to restrict β_i to positive

values, we set $\boldsymbol{b_i} \sim \mathcal{N}(0, C)$ and $d(a_i, \beta, \boldsymbol{b_i}) = \exp(\beta + \boldsymbol{b_i})$ (exponentiation here is applied component-wise). We won't consider α_i because our data doesn't provide measures of other cell-specific features. Let us define the vector of parameters $\beta_i = (c_{1,i}, c_{2,i}, c_{3,i}, c_{4,i}, c_{5,i}, c_{6,i}, c_{7,i}, c_{8,i}, x_{3(0),i}, \tau_i)$ and let $x_{6,i}(t|\beta_i)$ denote the state of an individual following the dynamics in (11) with parameters β_i, and initial conditions $x_{1(0),i} = 1$, $x_{2(0),i} = x_{4(0),i} = x_{5(0),i} = x_{6(0),i} = 0$. We can equip the system (11) with the output equation

$$y_i(t) = x_{6,i}(t|\beta_i) + \boldsymbol{e_i}(t) \tag{12}$$

where

$$\boldsymbol{e_i}(t) = (e_a + e_b x_{6,i}(t|\beta_i))\boldsymbol{\eta_i}(t). \tag{13}$$

Note that (12) and (13) are equivalent to (5) and (6). In particular, $\boldsymbol{\eta_i}(t), i = 1, .., n$ are independent white gaussian noises with unit intensity, and e_a^2,e_b^2, define the intensity of the additive and multiplicative parts of the measurement noise $e_{i,j}$. Then, in (2) we have $f(Z_{ij}, \beta_i) = x_{6,i}(t|\beta_i)$ and $e_{ij} = e_i(t_{ij})$. Regressors are $Z_{ij} = (t_{ij}, u(t)|_{t=0}^{t_{ij}})$ and the full set of parameters to be inferred is $\theta = \{\beta, C, e_a, e_b\}$. Model (13) is in agreement with the types of noise present in fluorescence microscopy [15]. Finally, to find the parameters in function (1) that maximize the marginal likelihood of the simulated distribution, we use the SAEM algorithm [16]. Starting values for β that provide good convergence properties are estimated as proposed in section 8.1 of [6], by calculating individual estimates first and then performing analysis of variability.

3) Model for MB inference: We chose the model proposed in [5] given the similarity of the systems and the promising results presented. We will use a zero-cumulant moment closure of the 35 equations reported in section S.4.1 of [5], which describe the time evolution of the population statistics. In this system of equations, the first 7 state variables represent the first-order moment (the mean) of each chemical species (including ribosomes) and the remaining equations represent the second-order uncentered moments (from which covariance between pairs of species can be inferred). We are interested in the first- and second-order moments of the fluorescent protein, which is denoted in [5] as the G species. The $c_k, k = 1, ..., 8$ parameters appearing in [5] will have the same meaning as the c_k parameters introduced in section III-B.2, with the exception of parameter c_1, which in our case acts as a scaling factor for the input u, and parameter c_6, which for simplicity of notation we have lumped together with the concentration of ribosomes (α_2^1 in [5]), given that they are mutually unidentifiable. We have also assumed the same sources of extrinsic and intrinsic variability as in [5]. The parameters to identify are a set of fixed parameters $\beta_f = (c_1, c_2, c_3, c_4, c_5, c_7, c_8, \tau)$ that are common to all cells, and statistics of a distribution on parameters $\beta_v = (x_3(0), c_6)$. In particular, we will identify average values $E[x_3(0)]$, $E[c_6]$ and the entries of the matrix $C_v = \text{cov}(x_3(0), c_6)$.
Extrinsic variability comes from the variable parameters and intrinsic variability results from the stochasticity of the CME.

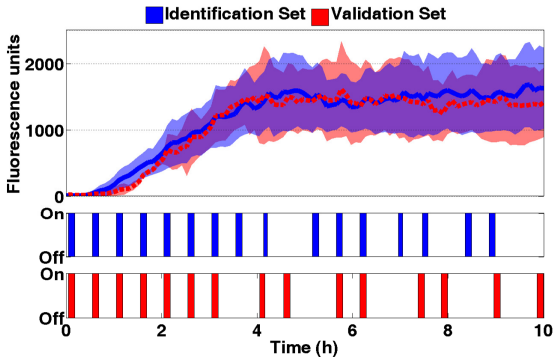

Fig. 2. Experimental data. The solid blue line and shaded blue area denote the mean +/- two standard deviations of the samples in the identification set. The dashed red line and shaded red area denote the same quantities of the validation sets. The pulses in the bottom of the figure represent the position of the valve that regulates the osmotic shocks applied to the system.

For the function $D(\cdot||\cdot)$ in (9) we will use the Kullback-Leibler divergence between Gaussian distributions. This is a simpler approximation than the one used in [5], but still provides a convenient (pseudo)metric for matching first- and second-order moments even if the distributions are not Gaussian. At every time instant j we will compute

$$D\big(\hat{m}_{y,j}, \hat{M}_{y,j}||m_{y,j}^{\theta}, M_{y,j}^{\theta}\big) = log(\frac{\sigma_{y,j}^{\theta}}{\hat{\sigma}_{y,j}}) + \frac{\hat{\sigma}_{y,j}^2 + (\hat{m}_{y,j} - m_{y,j})^2}{2\sigma_{y,j}^2} + \frac{1}{2} \tag{14}$$

$\hat{m}_{y,j}$, $m_{y,j}$ represent the mean of the observed and simulated distributions respectively, and $\hat{\sigma}_{y,j}$,$\sigma_{y,j}$ the corresponding standard deviations. Note that these values include the contribution of the measurement noise, as specified in (8). Starting values for β_f and β_v, are estimated in the same way as for ME inference.

C. Experimental data and data pre-processing

The data used consisted of a set of two experiments, in which cells had been controlled to follow a constant yECitrine concentration, set at 1500 fluorescence units. In each experiment, y_{ij} corresponds to the value of fluorescence for yECitrine in cell i at time t_j. The sampling time is 6 minutes and the analysis covers 10 hours of experiments. The empirical first and second moments of the data,

$$\hat{m}_{y,j} = \frac{1}{N}\sum_{i=1}^{N} y_{ij}\ , \qquad \hat{M}_{y,j} = \frac{1}{N}\sum_{i=1}^{N} y_{ij}^2,$$

correspond to the first-order and second-order uncentered moments of the yECitrine species in the MB model.
Raw data was processed as follows: first, a manual review of the snapshots taken in the two experiments was carried out to spot possible errors in tracking. This is because increasingly growing population and artifacts in the image capturing process led to misidentification of some cells during the real-time automatic analysis. At each time instant, cells whose fluorescence level was considered as outlier where marked, and cells marked as outliers more than 30% of their

lifetime were removed from the dataset. Finally, we set a time threshold based on the approximate lifespan of yeast cells [17], and only cells that where alive and successfully tracked for more than 90 minutes were included in the identification dataset.In Fig. 2, the identification and validation sets are compared. It can be seen that the mean values are very close to each other, as they represent the controlled outputs. The dispersion of the populations (represented in Fig. 2 by +/- two standard deviations) differs significantly, due to the difference in the initial number of cells. Note also that the input patterns are equal during the first 3 hours until the cells reach the desired output, but afterwards they differ.

IV. Inference results

A. Evaluation criteria

Time series of yECitrine concentrations for 10000 cells were computed using the ME and CME models. To discuss the results of the identification process, we assessed how well the identified models can reproduce identification and validation data. In the first case, we generated two datasets (ME1, MB1) using the input profile of the identification dataset (Figure 2, blue input). In the second case, we generated datasets ME2 and MB2 using the red input profile in Figure 2. Results are described in detail in sections IV-B and IV-C.

As explained in section III-B.2, in the ME case a cell is instantiated by extracting a random vector $\boldsymbol{b_i}$ from the distribution $\mathcal{N}(0, C)$, computing $\beta_i = \exp(\beta + \boldsymbol{b_i})$, and generating yEcitrine samples according to (12). For the CME, yECitrine samples for a cell are obtained by extracting a random vector $\beta_{v,i}$ from a lognormal distribution whose statistics are inferred from β_v, C_v , and then running Gillespie's algorithm [18] to sample from the probability $p(x, t)$ of the CME, instantiated with vectors $\beta_{v,i}$ and β_f. Finally, measurement noise is added according to (5). Identified models are then scored according to two criteria: the first one is the ability of the system to follow the time evolution of the mean and the standard deviation of the population. This was quantified in terms of the normalized RMSE

$$\text{NRMSE} = \frac{\sqrt{\frac{1}{T}\sum_{j=1}^{T}(\lambda_j - \hat{\lambda}_j)^2}}{\hat{\lambda}_{max} - \hat{\lambda}_{min}} \tag{15}$$

where, for an experiment spanning T time samples, $\hat{\lambda}_j$ is the jth sample of the analyzed feature (i.e. mean or standard deviation of the cell population) calculated from the observations, λ_j is the same quantity calculated from the simulated population, and $\hat{\lambda}_{max}$, $\hat{\lambda}_{min}$ are the maximum and minimum in the set $\{\hat{\lambda}\}_{j=1}^{T}$. In the sequel, NRMSE_M corresponds to the NRMSE of the mean trajectory and NRMSE_S is the NRMSE of the standard deviation of trajectories. The second criterion used is the Kolmogorov-Smirnov Two-Sample Test (KS), described in [19] and used to assess in a non-parametric way (i.e. without assumptions of the underlying continuous distribution) the similarity between two distributions. At each time instant, we compare yECitrine distribution over the simulated and real populations and a p-value (pK) is computed. The KS test is based on the null hypothesis that the samples are drawn from the same distribution and the null hypothesis is rejected with 95% confidence when pK$<$0.05, hence, a higher pK indicates higher similarity between the distributions. The percentage of time samples with pK$>$0.05 (i.e. the success rate), will be denoted with hK.

B. Performance on identification data

The computation time necessary to estimate ME and MB parameters, using an Intel(R) Core(TM) i7 CPU @2.67 GHz with 4 GB RAM, was in the order of 22 hours for ME (using nlmefitsa, from statistics toolbox [20]) and 1.5 hours (using fminsearch, from optimization toolbox [21]). In both cases only one processor was used. The huge difference in identification times is due to the fact that solving the optimization problem (9) requires to solve several times only the ODE system that characterizes the first and second moments of the population distribution; on the other hand, the minimization of the marginal likelihood in ME using the SAEM method, requires to simulate several times a whole population of cells using the model (11). Table 1 summarizes the results obtained when evaluating the model against the identification dataset. The NRMSE_M shows that MB inference has a slightly better performance following the mean. The standard deviation is also better followed by MB inference with a NRMSE_S significantly lower value for MB than for ME. The same general conclusion can be drawn from the KS test: the average of the pK value at all time instants is higher for MB as well as the hK success rate.

Table 1. Performance indices on identification data

	ME1	MB1
NRMSE_M	0.06	**0.04**
NRMSE_S	0.25	**0.11**
Avg. pK	0.25	**0.49**
hK	79%	**87%**

C. Model validation

Figure 3A reveals that the tracking of the mean is very good in both cases, but MB is better than ME in following the standard deviation, as the values in Table 2 confirm. Results in Table 2 show that quantitative differences between the two methods are overall small, although indicators NRMSE_M and NRMSE_S show a better performance using MB. The KS-test results don't favor MB in the validation stage as they did on the identification stage. Overall, ME presents an increase in performance between the identification and the validation stages, while MB's performance decreases for most indicators. This could be a sign that MB overfits the system, losing generalization capabilities. This aspect should be analyzed in detail in a future study in order to find criteria to prevent overfitting when using this technique.

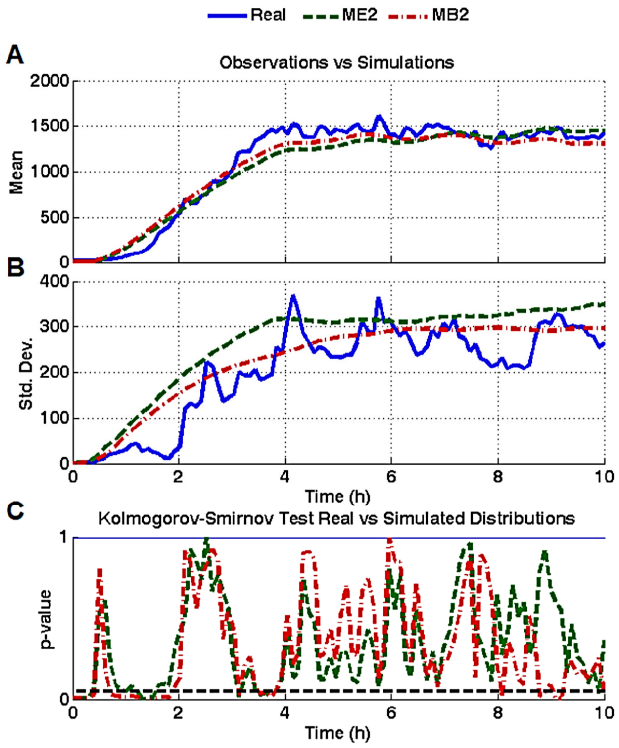

Fig. 3. Validation Experiments. Simulations are compared to data from a validation set. Figure (A) and (B) show the time evolution of ME and MB's identified models compared to the time evolution of the observations. (C) shows the time evolution of the pK value for both models, indicating how close the simulated distributions are to the real observed distributions at every time instant. The black dashed line represents the 95% confidence interval, below which, there is significant statistical evidence that the distributions are not equal.

Table 2. Performance indices on validation data

	ME2	MB2
$NRMSE_M$	0.08	**0.06**
$NRMSE_S$	0.20	**0.13**
Avg. pK	**0.34**	0.32
hK	**87%**	74%

V. Discussion and Concluding Remarks

In our experiments, both MB and ME have led to satisfactory results despite the differences in the frameworks and the assumptions on extrinsic or intrinsic variability. Therefore, the choice of the modeling framework depends on the particular problem addressed. Two factors that would favor the choice of ME are the built-in ability to account for expression profiles in individual cells, and the possibility of using characteristics such as cell size, cell-cycle state, and others, as covariates (vector α_i), to better quantify the deterministic factors that contribute to cell-to-cell variability [1].

One important aspect that can be only handled through single-cell data and therefore is much better suited for a ME approach, is the analysis of correlations in individual cell behaviors as a function of their parental relationships via cell lineage reconstruction. This can provide a better understanding of various cellular processes, such as aging [22] or cell cycle duration [23]. Because it allows to associate individual behaviors with individual parameter values, the ME framework seems ideally suited to model the effect of parameter correlations across many cell generations.

References

[1] B. Snijder and L. Pelkmans, "Origins of regulated cell-to-cell variability," *Nature Reviews, Molecular Cell Biology*, vol. 12, pp. 119–125, Feb. 2011.

[2] G. Balázsi, A. van Oudenaarden, and J. J. Collins, "Cellular Decision Making and Biological Noise: From Microbes to Mammals," *Cell*, vol. 144, pp. 910–925, Mar. 2011.

[3] B. Munsky, B. Trinh, and M. Khammash, "Listening to the noise: random fluctuations reveal gene network parameters," *Molecular Systems Biology*, vol. 5, Oct. 2009.

[4] J. P. Hespanha, "Moment closure for biochemical networks," in *Proceedings of the Third International Symposium on Control, Communications and Signal Processing*, March 2008.

[5] C. Zechner, J. Ruess, P. Krenn, S. Pelet, M. Peter, J. Lygeros, and H. Koeppl, "Moment-based inference predicts bimodality in transient gene expression," *Proceedings of the National Academy of Sciences*, vol. 109, pp. 8340–8345, May 2012.

[6] J. C. Pinheiro and D. M. Bates, *Mixed-Effects Models in S and S-PLUS*. Springer Verlag, New York, first ed., 2000.

[7] M. Davidian and D. M. Giltinan, *Nonlinear models for repeated measurement data*. Chapman & Hall, 1995.

[8] J. Uhlendorf, A. Miermont, T. Delaveau, G. Charvin, F. Fages, S. Bottani, G. Batt, and P. Hersen, "Long-term model predictive control of gene expression at the population and single-cell levels," *Proceedings of the National Academy of Sciences*, vol. 109, pp. 14271–14276, Aug. 2012.

[9] H. E. Samad, M. Khammash, L. Petzold, and D. Gillespie, "Stochastic modelling of gene regulatory networks," *International Journal of Robust and Nonlinear Control*, 2002.

[10] D. Wilkinson, *Stochastic Modelling for Systems Biology*. London, UK: Chapman & Hall/CRC, 2006.

[11] 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 Bioinformatics*, vol. 12, p. 125, Apr. 2011.

[12] S. Hohmann, "Osmotic Stress Signaling and Osmoadaptation in Yeasts," *Microbiology and Molecular Biology Reviews*, vol. 66, pp. 300–372, June 2002.

[13] H. de Jong, "Modeling and simulation of genetic regulatory systems: a literature review.," *Journal of Computational Biology*, vol. 9, pp. 67–103, Jan. 2002.

[14] A. M. Gonzalez, J. Uhlendorf, J. Schaul, E. Cinquemani, G. Batt, and G. Ferrari-Trecate, "Identification of biological models from single-cell data: a comparison between mixed-effects and moment-based inference," Tech. Rep. RR-8288, INRIA, April 2013.

[15] A. Jezierska, H. Talbot, C. Chaux, J.-C. Pesquet, and G. Engler, "Poisson-Gaussian noise parameter estimation in fluorescence microscopy imaging.," in *9th IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1663–1666, IEEE, May 2012.

[16] B. Delyon, M. Lavielle, and E. Moulines, "Convergence of a Stochastic Approximation Version of the EM Algorithm," *The Annals of Statistics*, vol. 27, no. 1, pp. 94–128, 1999.

[17] I. Herskowitz, "Life cycle of the budding yeast Saccharomyces cerevisiae.," *Microbiological Reviews*, vol. 52, pp. 536–553, Dec. 1988.

[18] D. T. Gillespie, "Exact stochastic simulation of coupled chemical reactions," *Journal of Phyical Chemistry*, vol. 81, pp. 2340–2361, December 1977.

[19] M. Hollander and D. A. Wolfe, *Nonparametric Statistical Methods, 2nd Edition*. Wiley-Interscience, 2 ed., Jan. 1999.

[20] The MathWorks, "Matlab statistics toolbox," R2011a.

[21] The MathWorks, "Matlab optimization toolbox," R2011a.

[22] A. B. Lindner, R. Madden, A. Demarez, E. J. Stewart, and F. Taddei, "Asymmetric segregation of protein aggregates is associated with cellular aging and rejuvenation," *Proceedings of the National Academy of Sciences*, vol. 105, pp. 3076–3081, Feb. 2008.

[23] G. Charvin, F. R. Cross, and E. D. Siggia, "A microfluidic device for temporally controlled gene expression and long-term fluorescent imaging in unperturbed dividing yeast cells," *PLoS One*, vol. 3, p. e1468, January 2008.