Validation Methods for Population Models of Gene Expression Dynamics

Andres M. Gonzalez-Vargas, Eugenio Cinquemani, Giancarlo Ferrari-Trecate

Validation methods for population models of gene expression dynamics

Andrés M. González-Vargas [*] **Eugenio Cinquemani** [**] **Giancarlo Ferrari-Trecate** [***,****]

[*] *Departamento de Automática y Electrónica. Universidad Autónoma de Occidente. Cll 25#115-85 Km 2 Vía Cali-Jamundi. Cali, Colombia (e-mail: amgonzalezv@uao.edu.co)*
[**] *INRIA Grenoble – Rhône-Alpes, Montbonnot, 38334 St.Ismier Cedex, France (e-mail: eugenio.cinquemani@inria.fr)*
[***] *Dipartimento di Ingegneria Industriale e dell'Informazione, Università degli Studi di Pavia, via Ferrata 3, 27100 Pavia, Italy (e-mail: giancarlo.ferrari@unipv.it).*
[****] *Currently: Automatic Control Laboratory, École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland*

Abstract: The advent of experimental techniques for the time-course monitoring of gene expression at the single-cell level has paved the way to the model-based study of gene expression variability within- an across-cells. A number of approaches to the inference of models accounting for variability of gene expression over isogenic cell populations have been developed and applied to real-world scenarios. The development of a systematic approach for the validation of population models is however lagging behind, and accuracy of the models obtained is often assessed on a semi-empirical basis. In this paper we study the problem of validating models of gene network dynamics for cell populations, providing statistical tools for qualitative and quantitative model validation and comparison, and guidelines for their application and interpretation based on a real biological case study.

Keywords: Statistical methods, System Biology, Stochastic modelling, Mixed-Effects modelling, Gene expression

1. INTRODUCTION

Modern experimental techniques for the monitoring of gene expression at the individual cell level provide both evidence of gene expression variability, and quantitative data that can be exploited to describe and analyze variability from a mathematical standpoint (Elowitz et al., 2002; Neuert et al., 2013). Various approaches to the modelling of gene expression variability within and across cells have been developed, along with methods for their inference from experimental data, and applied to the study of real biological systems (Munsky et al., 2009; Zechner et al., 2012; Neuert et al., 2013; Llamosi et al., 2016). Yet, the quality of these models is often difficult to assess, due to the inherent complexity of the models as well as the challenges and costs involved in conducting validation experiments. Model assessment is mostly performed on empirical bases, such as qualitative response shape (Munsky et al., 2009; Zechner et al., 2012), overexpression or knock-out experiments (Cantone et al., 2009), and so on, whereas quantitative predictive capabilities are largely unexplored. The aim of this work is to introduce systematic approaches for the validation of mathematical models of cellular response variability. We are interested in particular in population modelling, i.e. the ability to account for response variability across different cells. Validation methods that will be considered shall emphasize the predictive capabilities of the models, i.e. the ability to correctly anticipate the true system response in new and possibly different experimental conditions. For parametric models, in particular, this rules out approaches based on the analysis of estimated parameter, because parameter inaccuracies are hardly related with predictive capabilities in the common scenario where practical identifiability issues arise (Gutenkunst et al., 2007). For practical utility, methods should be applicable with no further effort by modellers. We will therefore restrict to general validation tools, avoiding to leverage specificities of the different modelling approaches.

We will start by reviewing Mixed-Effects (Lavielle, 2015) and Chemical Master Equation (El Samad et al., 2005) modelling, two somewhat complementary approaches to population modelling that represent well the variety of modelling approaches currently proposed in the literature. We will also summarize the more traditional Mean-Cell modelling, for comparison purposes. Based on simulation of a biological case study, we will infer these models from *in silico* generated data and use them as a running example to introduce and discuss several validation methods derived from the statistical literature. We will illustrate their application for the evaluation of individual models as well as for model comparison, showing that reliable conclusions can be drawn from the ensemble of validation results rather than from the application of a single tool.

2. POPULATION MODELS FOR GENE EXPRESSION DYNAMICS

Gene expression dynamics are generally given in terms of a biochemical reaction network operating in a uniform volume, a convenient abstraction of a cell (or a portion of it, e.g. the nucleus). Such a network is then simply characterized by n species, m reaction channels, and a stoichiometry matrix ν with n rows and m columns, each column describing the net change in copy number of the n molecular species over the whole reaction volume when the corresponding reaction takes place. Let $x = [x_1, \dots, x_n]^T$ denote the amount of molecules of every species. Network dynamics are then fixed by the reaction rates $v(x, \psi)$, an m-dimensional column vector whose entries quantify the velocity at which different reactions take place. As apparent from the notation, $v(x, \psi)$ generally depends on the amount of molecules present in the reaction volume, and on kinetic rate parameters that are typically unknown or only partially known, and need to be determined from experimental data. In more generality, reactions may depend on (possibly time-varying) exogenous variables u affecting rates (e.g. a control signal), in which case we write $v(x, u, \psi)$.

Population models aim at applying this general paradigm to the description of multiple entities (cells) that, despite identical in principle and hence obeying the same model structure, show different responses. Several approaches may be considered, further detailing the meaning of x, ψ and v, as reviewed below.

Mean-Cell (MC) modelling. This approach aims at describing some "typical" behavior of a cell. For a given species abundance x_0 at a time t_0, a deterministic response model for the abundances $x(t)$ at all times t is sought. Under appropriate assumptions on reaction volume and species abundance, allowing in particular to treat $x(t)$ as species concentrations, the entries of $v(x, u, \psi)$ admit the interpretation of (deterministic) number of reaction occurrences per unit time, and are determined by the laws of mass action. In addition, $x(t)$ obeys

$$\dot{x}(t) = \nu v\big(x(t), u(t), \psi\big) \tag{1}$$

with $x(t_0) = x_0$. When confronted with population-average data, x is interpreted as a vector of average concentrations across the cell population, and ψ are considered as typical kinetic parameters. In the context of population modelling, where single-cell profiles are generally different from one another, the solution of (1) is rather interpreted as "mean-cell" dynamics, an oversimplification of the ensemble of single-cell responses. Single-cell measurements are then described as

$$y_i(t) = f\big(t, u(\cdot), x_0, \psi\big) + \text{error}_i$$

where f is determined by the solution of the above ODE for given parameters ψ and initial conditions x_0 under $u(\cdot)$, while error_i accounts for the discrepancy between the mean-cell response f and the response of the ith of N cells, as well as for measurement noise.

Inference of MC models can be addressed by Maximum Likelihood (ML). Suppose that, for every cell $i = 1, \dots, N$, measurements $\mathcal{Y}_i = \{y_{i,j} = y_i(t_j) : \ j = 1, \dots, T_i\}$ are collected at times $\mathcal{T}_i = \{t_{i,j} : \ j = 1, \dots, T_i\}$, and denote with $\mathcal{Y}$ the complete dataset. Consider a generic measurement model of the type

$$y_{i,j} = f\big(t_j, u(\cdot), x_0, \psi\big) + h\big(f\big(t_j, u(\cdot), x_0, \psi\big), \epsilon\big)\eta_i(t_j) \tag{2}$$

where errors $\eta_i(t_j) \sim \mathcal{N}(0, 1)$ are mutually independent across i and j, and ϵ are parameters of the noise distribution. Note that h plays the role of error standard deviation, which may be affected in different ways from the current system state. Denoting $\theta = (\psi, \epsilon)$ the set of unknown parameters (possibly including x_0), the ML estimate of θ may be computed by minimizing its negative log-likelihood given $\mathcal{Y}$, i.e., for $f_j(\theta) = f\big(t_j, u(\cdot), x_0, \psi\big)$ and $h_j(\theta) = h\big(f_j(\theta), \epsilon\big)$,

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{N} \sum_{j=1}^{\mathcal{T}_i} \left\{ \frac{1}{2} \left(\frac{y_{i,j} - f_j(\theta)}{h_j(\theta)} \right)^2 + \log h_j(\theta) \right\}.$$

Mixed-Effects (ME) modelling. An alternative approach is to assume that (1) models the individual cell, but different cells may be characterized by different values of ψ. If ψ_i denotes the parameters of the ith cell, one then assumes that

$$y_i(t) = f\big(t, u(\cdot), x_0, \psi_i\big) + \text{error}_i, \quad \text{(individuals model)}$$

where $f\big(t, u(\cdot), x_0, \psi_i\big)$ is the solution of (1) with $\psi = \psi_i$, and error_i accounts for the inaccuracy in modelling single-cell response (and measurement noise). Here x is thought of as concentrations in the relevant cell, and $v(x, u, \psi_i)$ the velocity of reactions in cell i for given intracellular concentrations, while u is common across the population. Mixed-effects modelling enforces the idea of a cell being a variant of a statistically homogeneous population by introducing a common prior on parameters ψ_i,

$$\psi_i = d(a_i, \mu, b_i), \ b_i \sim \mathcal{N}(0, \Omega), \quad \text{(population model)}.$$

The entries of the parameter vector μ, common to the whole population, are called fixed-effects. Vectors b_i are mutually independent and contain the random effects, i.e. individual cell discrepancies from the population average. Finally a_i are covariates representing cell-specific known features, if present.

Inference of mixed-effects models from individual data has the primary aim of reconstructing the population properties μ and Ω from the whole dataset $\mathcal{Y}$ of all measurements from all individuals. Consider again a generic measurement model of the form (2), where ϵ is fixed across individuals and ψ is replaced by ψ_i. A statistically powerful approach is provided by Population Likelihood Maximization (PLM). The idea is to leverage all data $\mathcal{Y}$ at once by maximizing with respect to $\Theta = (\mu, \Omega, \epsilon)$ the marginal likelihood $p(\mathcal{Y}|\Theta) = \prod_{i=1}^{N} \int d\psi_i p(\mathcal{Y}_i|\psi_i, \epsilon) p(\psi_i|\mu, \Omega)$, where factorization occurs thanks to the mutual independence of the b_i and of the η_i. By this approach, a single estimate is obtained for all population parameters, including ϵ. From the resulting estimates $\hat{\mu}$ and $\hat{\Omega}$, single-cell parameter estimates $\hat{\psi}_i$ may also be computed, e.g., by maximizing the empirical posterior $p(\psi_i|\mu = \hat{\mu}, \Omega = \hat{\Omega})$. In practice, while all integrands can be written explicitly, no closed form expression exists in general for $p(\mathcal{Y}|\Theta)$. Numerical methods for approximate PLM have been proposed (notably NONMEM (Bauer et al., 2007) and SAEM (Delyon et al., 1999; Bauer et al., 2007)) and are contained in dedicated software packages such as Monolix (Lixoft, 2014).

Chemical Master Equation (CME) modelling. Models above rely on deterministic dynamics for single cells. This is inadequate when randomness of gene expression and regulation is prominent. At the single-cell level, gene expression noise can be captured by modelling the process as a (stochastic) Markov Chain. Let x be a count of molecules of the different species, and interpret $v(x,\psi)$ as infinitesimal probabilities that the different reactions occur in an infinitesimal period of time. These rates are typically determined by mass-action laws, and ψ are the corresponding kinetic constants (Gillespie, 1992). As a result, $x(t)$ obeys the laws of a Markov process (possibly driven by a control input u). For all possible values z of $x(t)$, the probabilities $p^{\psi}(z,t) = \text{Prob}\big(x(t)=z\big|\psi\big)$ evolve over time in accordance with an infinite-dimensional ODE called CME (see e.g. El Samad et al. (2005)). In sharp contrast with ME modeling, the underlying assumption is that the same model with identical parameters ψ applies to all cells, so that different cell profiles are different outcomes of the same stochastic process. Mixtures of ME and CME models have also been proposed (Zechner et al., 2014), but we will not address them here.

In the current literature, CME models are mostly inferred from empirical statistics of $z(t)$ computed from independent cell samples at different time points t (Munsky et al., 2009; Zechner et al., 2012). Measurements $\tilde{y}(t_j)$ at time points $\mathcal{T} = \{t_j : \ j = 1,\dots,T\}$ can thus be seen as a noisy readout of $p^{\psi}(\cdot,t_j)$. The task is to estimate ψ from $\tilde{\mathcal{Y}} = \{\tilde{y}_j : \ t_j \in \mathcal{T}\}$ ("~" is used here to distinguish measurements of statistics of x from measurements of x itself). Here we consider an approach known as Moment Matching (MM). Solutions based on the approximation of the CME also exist (Munsky et al., 2009; El Samad et al., 2005). Let $M^{\psi}(t)$ be the vector containing the moments of $x(t)$ up to order L. It can be shown (Schnoerr et al., 2015) that $\dot{M}^{\psi}(t) = A(\psi)M^{\psi}(t) + B(\psi)\bar{M}^{\psi}(t)$ for some matrices A and B depending on the network reaction rates (and ν), where $\bar{M}^{\psi}(t)$ denotes moments of order higher than L. The equation for $M^{\psi}(t)$ is generally "open" ($B \neq 0$), i.e. it cannot be solved due to the unknown moments $\bar{M}^{\psi}(t)$. However, several so-called moment closure methods have been proposed, resulting in "closed" systems of equations

$$\dot{\tilde{M}}^{\psi}(t) = A(\psi)\tilde{M}^{\psi}(t) + \phi\big(\tilde{M}^{\psi}(t)\big) \tag{3}$$

whose solution $\tilde{M}^{\psi}$ approximates M^{ψ} in a way that depends on the definition of ϕ (Schnoerr et al., 2015; Zechner et al., 2012). Measurements then obey

$$\tilde{y}_j = c^T\tilde{M}^{\psi}(t_j) + h\big(\tilde{M}^{\psi}(t_j),\epsilon\big)\eta(t_j), \tag{4}$$

with usual assumptions on η, and vector c accounting for partial observation of $\tilde{M}^{\psi}$. In particular, if $L = 2$, then (3) involves mean, variance and covariance terms, whereas only mean and variance for a single species are provided by most common experimental setups, such as the one illustrated in this paper. Inference then becomes finding the value of ψ that best explains (4), with $\tilde{M}^{\psi}$ the solution of (3). Different solutions can be found depending on the specific characterization of h (Zechner et al., 2012; Gonzalez et al., 2013). Here we will apply the method in Gonzalez et al. (2013), where an additive-multiplicative noise model $h\big(f,\epsilon\big) = \epsilon_a + \epsilon_b f$ is assumed, with parameters ϵ_a and ϵ_b also estimated from the data.

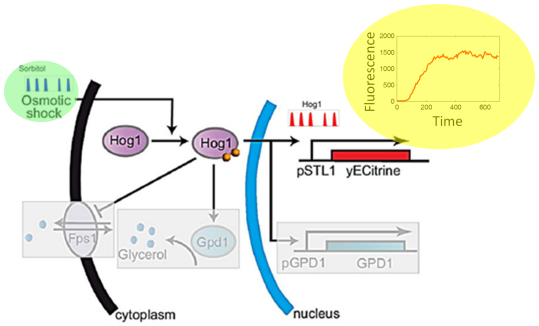

Fig. 1. Hyperosmotic gene expression in yeast. Hyperosmotic stress triggers phosphorylation and nuclear import of the protein Hog1, which thereupon activates osmo-stress responsive genes. In our reference setup (Uhlendorf et al., 2012; Llamosi et al., 2016), a fluorescent reporter gene sequence (yECitrine) is engineered in the cells under the control of osmosensitive promoter pSTL1, which results in the synthesis of fluorescent reporter molecules upon cell sensing of osmotic shocks. Additional response and adaptation mechanisms (shaded in gray) are prevented in the specific experimental setup, and will not be considered here.

2.1 Example: Yeast osmotic shock response

In order to discuss validation methods for population models inferred from biological data, we will consider the case study of osmotic shock response in yeast *Saccharomices cerevisiae* cells (Llamosi et al., 2016). The biological system is illustrated in Fig. 1. We will only be concerned with the modelling of the expression of the reporter gene as a function of the osmolarity shocks delivered to yeast cells by means of a computer-controlled microfluidics system (see details in Uhlendorf et al. (2012)). Perception of an osmotic shock (u_h) leads to the activation of the osmosensitive genes, resulting in particular in the transcription of fluorescent reporter mRNA molecules (mRNA), subsequently translated into immature protein molecules ($\text{Protein}^{\text{off}}$). A subsequent maturation step takes proteins in their mature, fluorescent form ($\text{Protein}^{\text{on}}$). All species are also subject to degradation and dilution due to cell growth. In accordance with Gonzalez et al. (2013), the system is then represented by the following reactions:

$$\emptyset \xrightarrow{k_5 u_h} \text{mRNA} \qquad \text{mRNA} \xrightarrow{k_6} \emptyset \tag{5}$$

$$\text{mRNA} \xrightarrow{k_7} \text{mRNA} + \text{Protein}^{\text{off}} \qquad \text{Protein}^{\text{off}} \xrightarrow{k_8} \emptyset \tag{6}$$

$$\text{Protein}^{\text{off}} \xrightarrow{k_9} \text{Protein}^{\text{on}} \qquad \text{Protein}^{\text{on}} \xrightarrow{k_8} \emptyset \tag{7}$$

where the indexing of reaction rate constants is chosen for consistency with the same work. In turn, the shock perceived by cells u_h is related with the concentration u_c of a chemical inducer in the microfluidics chambers where the cells reside via the equation $\dot{u}_h(t) = k_h u_c(t) - \gamma_h u_h(t)$. Quantity u_c represents the concentration manipulated by the experimenter, i.e. the system input previously called u. Via an automatic microscopy image acquisition and processing system, measurements of cell fluorescence, i.e. the concentration of $\text{Protein}^{\text{on}}$, are collected over time. A full characterization of the experimental platform is provided in Uhlendorf et al. (2012). For mean-cell and ME modelling, denoting with $x = [x_1, x_2, x_3]^T$ the concentrations of mRNA, $\text{Protein}^{\text{off}}$ and $\text{Protein}^{\text{on}}$, in the same order, after solving for the system stoichiometry and the mass-action reaction velocities we get that

$$\dot{x}_1(t) = k_5 u_h(t) - k_6 x_1(t), \quad (8)$$
$$\dot{x}_2(t) = k_7 x_1(t) - (k_8 + k_9) x_2(t), \quad (9)$$
$$\dot{x}_3(t) = k_9 x_2(t) - k_8 x_3(t). \quad (10)$$

For ME models, parameters $\psi_i = (k_5, k_6, k_7, k_8, k_9)$ are cell-dependent. For the ith cell, fluorescence measurements are considered to be of the form

$$y_i(t) = x_3(t) + \big(\epsilon_a + \epsilon_b x_3(t)\big)\eta_i(t). \quad (11)$$

We will use this model to generate data *in silico* and discuss validation of the various models described above.

3. VALIDATION OF CELL POPULATION MODELS

In this section we present validation criteria for assessing the quality of cell population models. Several approaches come from the literature on ME models (Pinheiro and Bates, 2000; Comets et al., 2010). Their application will be discussed using the biological example in Section 2.1. To this purpose, we simulate 100 cells using a ME model based on (8)–(11). The osmotic stress profile and single-cell profiles $y_i(t)$ are shown in Gonzalez-Vargas et al. (2016). Parameters $(k_5, k_6, k_7, k_8, k_9)$ are sampled from a multivariate log-normal distribution, whose mean and covariance matrix, in log-scale, are $\boldsymbol{\mu} = [3.40\ -1.22\ -0.05\ -5.52\ -4.04]$, $\boldsymbol{\Omega} = 0.1\mathbf{I}_5$ (Gonzalez et al., 2013; Llamosi et al., 2016). We will infer three models (MC, CME and ME): the predictions of each model will be compared against the reference dataset, and we will show how validation methods can be useful to ascertain the model accuracy. Then, in Section 3.1 we will describe how the validation criteria can be jointly used for assessing acceptability of a model.

NRMSE and relative error. These two indicators are frequently used as a quantitative aid for the validation methods known as population plots (see later). The Root Mean Squared Error (RMSE) represents the sample standard deviation of the prediction error, i.e. the difference between predicted and observed values. As RMSE is scale-dependent, it is often common to normalize it in order to provide a scale-independent measure. The Normalized Root Mean Squared Error (NRMSE) is defined as

$$\mathrm{NRMSE}(\lambda, \hat{\lambda}) = \frac{1}{\lambda_{max} - \lambda_{min}} \cdot \sqrt{\frac{1}{T}\sum_{j=1}^{T}(\lambda_j - \hat{\lambda}_j)^2} \quad (12)$$

where, for an experiment spanning T time samples, λ_j is the j -th sample of the variable under analysis, e.g. a single-cell trajectory, the mean trajectory of the cell population, or the moments of the distribution of trajectories. The predicted values of the variable under study are $\hat{\lambda}_j$, and λ_{max}, λ_{min} are the maximum and minimum values in the full set of data. Furthermore, λ and $\hat{\lambda}$ in (12) denote the set of observed and predicted values, respectively.

Population plots. A simple way to compare the predicted and observed cell populations is to compute at every time instant j the empirical mean and standard deviation

$$\hat{m}_{\mathcal{Y},j} = \frac{1}{\mathrm{N}_j}\sum_{i\in\mathcal{N}_j}\mathcal{Y}_{ij},\ \hat{\sigma}_{\mathcal{Y},j} = \sqrt{\frac{1}{\mathrm{N}_j - 1}\sum_{i\in\mathcal{N}_j}(\mathcal{Y}_{ij} - \hat{m}_{\mathcal{Y},j})^2}$$

and compare them with the same quantities $(m_{y,j}, \sigma_{y,j})$ computed from a dataset of simulated cells created using the identified model. The observed and predicted statistics $(\hat{m}_{\mathcal{Y},j}, \hat{\sigma}_{\mathcal{Y},j})$, $(m_{y,j}, \sigma_{y,j})$ will then be used for plotting the mean together with a standard-deviation band in a single picture, called *standard plot*. Standard plots for the models of interest are shown in Gonzalez-Vargas et al. (2016). The standard plot provides information about the location and dispersion of the population (implicitly assuming Gaussianity of the underlying distributions), but it does not take into account single-cell fits.

Visual Predictive Check (VPC). VPC is a popular method for evaluating nonlinear ME models in population pharmacometrics (Comets et al., 2010; Lavielle, 2015). The idea behind the VPC is to assess graphically whether simulations from a proposed model are able to reproduce the central trend and variability in the measured data. The VPC does not make any assumption on the form of the distributions and also takes into account the uncertainty generated by calculating population statistics on small samples. The procedure uses the estimated model parameters and the design structure of the observed data, (input, time, and number of samples) to generate K datasets, each of N simulated cells. Then, in each dataset we compute the 0.5, 0.025 and 0.975 quantiles. Having K estimates of each quantile we can compute and plot a confidence interval for them, which makes the interpretation of VPCs less subjective. Finally, one can overlap "prediction bands" with estimated quantiles from the observed data. In this general form, the VPC provides a visual comparison of the overlap between the simulated distribution with the observations. Fig. 2 shows the classic VPC for the models of interest.

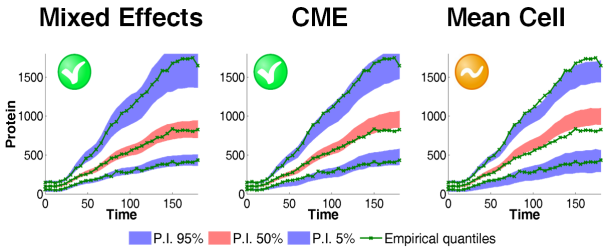

Fig. 2. VPC: shaded areas denote 99% confidence intervals on the calculated quantiles for the predicted dataset. The selected quantiles are 0.025 (blue), 0.5 (red) and 0.0975 (blue) which comprise 95% of the population. The green lines show the same quantiles for the reference dataset. A large deviation of the reference quantiles from the predicted quantiles' area suggests misspecification in the model.

Kolmogorov-Smirnov test. The Kolmogorov-Smirnov Two-Sample (KS2) test (Smirnov, 1939) is used to assess the similarity between two sample distributions without assumptions on the true distributions. In order to compute the KS statistic we generate a set of N' (typically $N' \geq 10000$) simulated cells using the identified model. We compute, at each time instant, $F_{1,N}(x)$ and $F_{2,N'}(x)$, which are, respectively, the Empirical Cumulative distribution Functions (ECDFs, see Gonzalez-Vargas et al. (2016); Rice (2006)) of the observed and simulated datasets. Then we compute $D_{o-p} = \sup_x |F_{1,N}(x) - F_{2,N'}(x)|$, where sup is the supremum function, and D_{o-p} is the distance between the two distributions. The test's null hypothesis is that both samples are drawn from the same distribution and this is

rejected at significance level $1-\alpha$ if $D_{o-p} > c(\alpha)\sqrt{\frac{N+N'}{NN'}}$. If we choose a significance level of 95% then $\alpha = 0.05$ and $c(\alpha) = 1.36$ (Miller, 1956). The result is given by a Boolean value hK equal to 1 if the null hypothesis is rejected and 0 otherwise. Based on this indicator, we can calculate a success rate S_{ks} for the whole experiment as $S_{ks} = 1 - \frac{1}{T}\sum_{j=1}^{T} hK_j$. We can also compute the average p-value of S_{ks}. A higher p-value will indicate that the two distributions are more similar (see Fig. 3).

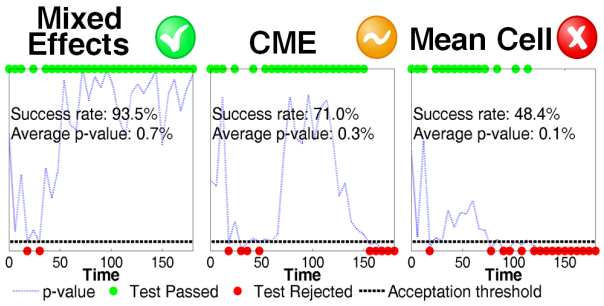

Fig. 3. KS2 test. The blue line represents the p-value obtained from the test at each time instant (the higher the better). The 95% threshold p-value (black-dashed line) separates unsuccessful time points (red points, indicating the distributions are statistically different) from successful time points (green).

Prediction distribution errors. The Prediction Distribution Errors (PDE) are proposed in Comets et al. (2010) as a metric to evaluate the performance of a ME model, based on Monte Carlo simulations of the population. We start by constructing a simulated dataset of K "repetitions" (i.e. cells simulated with the identified model) for each of the N observed cells. Ideally the number of repetitions should be as high as possible (usually $K \geq 1000$). Observations produced by the same individual at different time instants are correlated and the first step for deriving PDEs is to decorrelate them (see Gonzalez-Vargas et al. (2016) for details). Let us denote with $y_i^{\mathrm{sim}(k)*}$ the decorrelated vector of simulated observations for the ith cell in the kth simulation and with $\mathcal{Y}_i^*$ the decorrelated vector of real observations for the ith subject. Then, we can calculate the PDE as $\mathrm{PDE}_{ij} = \frac{1}{K}\sum_{k=1}^{K} \delta_{ijk}^*$, where $\delta_{ijk}^* = 1$ if $y_{ij}^{\mathrm{sim}(k)*} < \mathcal{Y}_{ij}^*$ and 0 otherwise. PDE values are (theoretically) decorrelated over time for the same individual and they follow a uniform distribution $\mathcal{U}(0,1)$ even when there are several observations per cell. A normalized version of PDE (NPDE) can be obtained by using the inverse function of the normal cumulative density function F: $\mathrm{NPDE}_{ij} = F^{-1}(\mathrm{PDE}_{ij})$.
Results of the NPDE can be seen in Fig. 4. In the top row, quantile-quantile plots give us a visual indication of how close the quantiles of NPDE overlap with those of a standard normal distribution. They should be as aligned as possible. The Bonferroni p-value included in the plot (see Gonzalez-Vargas et al. (2016) for a description) gives us a numerical indication of how close the distribution of the NPDE resembles a standard Gaussian distribution. The same comparison can be done using the plot in the bottom row in Fig. 4.

A posteriori best fits (APBFs). Using the simulated datasets introduced for discussing PDEs, we can compute,

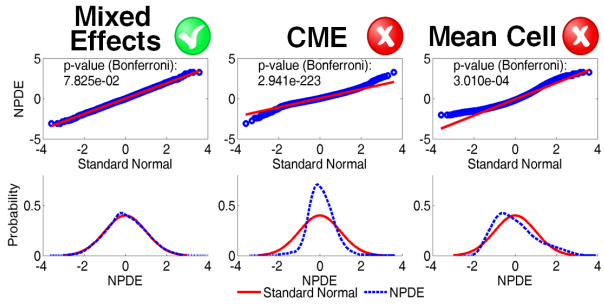

Fig. 4. Normalized prediction distribution errors (NPDE). Quantile-Quantile plots (top) compare the NPDE distribution (blue circles) to a normal standard distribution (red line). The Bonferroni-corrected p-value quantifies the closeness of both distributions. The bottom plots show the same comparisons, but from the perspective of probability density functions.

for observed cell i, $\mathrm{APBF}_i = \arg\min_k(\mathrm{NRMSE}(y_i^{sim(k)}, \mathcal{Y}_i))$. In other words, APBF_i denotes the index k that minimizes the NRMSE between $y_i^{\mathrm{sim}(k)}$ and $\mathcal{Y}_i$. Then, we can obtain a visual indication of the goodness of fit, by plotting best fits vs observations, and computing a numerical indicator of the total goodness of fit (i.e., for all cells): $\mathrm{NRMSE}_{\mathrm{APBF}} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{NRMSE}(y_i^{\mathrm{sim}(\mathrm{APBF}_i)}, \mathcal{Y}_i)$. When two models perform equally well at the population level, one can use APBF to choose which one performs better at the single-cell level. The lower the $\mathrm{NRMSE}_{\mathrm{APBF}}$, the better the model is able to represent individual cells. Fig. 5 shows APBF plots for the models of interest.

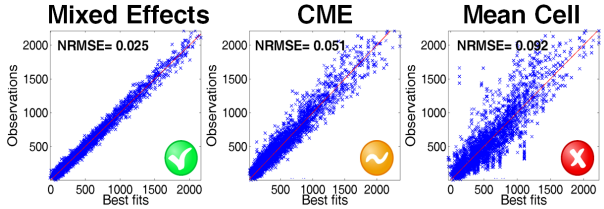

Fig. 5. APBF plots. Points $(y_{ij}^{\mathrm{sim}(\mathrm{APBF}_i)}, y_{ij})$ for $i = 1, \ldots, N$, $j = 1, \ldots, T_i$ are represented, i.e. individual best predictions against observed values of the reference data. A lower spread of the points in the anti-diagonal direction indicates better agreement between observations and predictions; this can be quantified by calculating $\mathrm{NRMSE}_{\mathrm{APBF}}$ (see text).

3.1 Joint use of validation criteria

Since different validation criteria are available for cell population models, in this section we provide guidelines for combining them so as to compare different models and, if possible, to isolate the best one, still with reference to the *in silico* results reported above.
As discussed in Section 3, several validation approaches require to simulate data using the identified models. To this purpose, we created datasets of 10000 cells each. Validation results are shown in Fig. 2–5. For an easier visual comparison between models, we display in all figures colored circles indicating good (green), moderate (yellow) and bad (red) results.
Based on standard plots (not reported here, see Fig. 3 of Gonzalez-Vargas et al. (2016)), all models seem to perform equally well. Another visual evaluation is provided by VPC (Fig. 2). The green lines representing the empirical

quantiles of the reference data tend to fall outside of the MC predicted quantiles. This gives us some preliminary evidence of model misspecification in the MC case.

A quantitative assessment is provided by the KS2 test. The success rates reported in Fig. 3 provide strong evidence to discard the MC model, and hint at possible inadequacy of CME modelling.

Different from the previous tests, focused on the models' ability to reproduce population statistics, the APBF method (Fig. 5) verifies the ability of a model to reproduce individual cell profiles by comparing each of the reference cells to the corresponding best-fitting cell from the predicted dataset. If the predicted model is able to fit sufficiently well the individual cells, all the blue crosses in Fig. 5 should be very close to the diagonal. The best model should show little dispersion in the anti-diagonal direction and the smallest NRMSE. The NRMSE indicates that ME is better than the two competing models.

Finally, the NPDE approach evaluates in some sense both individual cell and population performance of the models. The Q-Q and PDF plots in Fig. 4 show that the NPDE of the ME model follow very closely a standard normal distribution, while the CME and MC deviate from it noticeably. In summary, based on the last two tests we have a strong evidence in favor of the ME model, which corresponds to the actual model used to generate the reference dataset.

This example shows that simple visual checks of mean and standard deviation can give an erroneous idea of goodness of fit, which can be partially solved by using more complete indicators such as VPC.

4. CONCLUSIONS

In this paper, we have compared methods for validating models of cell populations. Existing validation approaches are still generic, in the sense that they can be applied to population of systems, even outside the context of Biology. As validation approaches can be useful for discriminating the relative importance of different sources of biological noise, future research will focus on incorporating genuine biological aspects in their formulation.

REFERENCES

Bauer, R.J., Guzy, S., and Ng, C. (2007). A survey of population analysis methods and software for complex pharmacokinetic and pharmacodynamic models with examples. *AAPS JOURNAL*, 9(1).

Cantone, I., Marucci, L., Iorio, F., Ricci, M.A., Belcastro, V., Bansal, M., Santini, S., di Bernardo, M., di Bernardo, D., and Cosma, M.P. (2009). A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. *Cell*, 137(1), 172 – 181.

Comets, E., Brendel, K., and Mentrè, F. (2010). Model evaluation in nonlinear mixed effect models, with applications to pharmacokinetics. *Journal de la Sociètè Francaise de Statistiques*, 151, 106–128.

Delyon, B., Lavielle, M., and Moulines, E. (1999). Convergence of a Stochastic Approximation Version of the EM Algorithm. *The Annals of Statistics*, 27(1), 94–128.

El Samad, H., Khammash, M., Petzold, L., and Gillespie, D. (2005). Stochastic modelling of gene regulatory networks. *International Journal of Robust and Nonlinear Control*, 15(15), 691–711.

Elowitz, M.B., Levine, A.J., Siggia, E.D., and Swain, P.S. (2002). Stochastic gene expression in a single cell. *Science*, 297(5584), 1183–1186.

Gillespie, D.T. (1992). A rigorous derivation of the chemical master equation. *Physica A*, 188(1), 404 – 425.

Gonzalez, A.M., Uhlendorf, J., Schaul, J., Cinquemani, E., Batt, G., and Ferrari-Trecate, G. (2013). Identification of biological models from single-cell data: a comparison between mixed-effects and moment-based inference. In *Proceedings of the 12th ECC*, 3652–3657.

Gonzalez-Vargas, A.M., Cinquemani, E., and Ferrari-Trecate, G. (2016). Validation methods for population models of gene expression dynamics. Research Report RR-8938, INRIA Grenoble - Rhône-Alpes. URL `https://hal.inria.fr/hal-01349030`.

Gutenkunst, R.N., Waterfall, J.J., Casey, F.P., Brown, K.S., Myers, C.R., and Sethna, J.P. (2007). Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput Biol*, 3(10), 1–8.

Lavielle, M. (2015). *Mixed-Effects models for the population approach*. CRC press.

Lixoft (2014). *Monolix User Manual Version 4.3.2*. Lixoft.

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), 1–18.

Miller, L.H. (1956). Table of Percentage Points of Kolmogorov Statistics. *Journal of the American Statistical Association*, 51(273).

Munsky, B., Trinh, B., and Khammash, M. (2009). Listening to the noise: random fluctuations reveal gene network parameters. *Molecular Systems Biology*, 5.

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

Pinheiro, J.C. and Bates, D.M. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer Verlag, New York.

Rice, J. (2006). *Mathematical statistics and data analysis*. Nelson Education.

Schnoerr, D., Sanguinetti, G., and Grima, R. (2015). Comparison of different moment-closure approximations for stochastic chemical kinetics. *The Journal of Chemical Physics*, 143(18), 185101.

Smirnov, N.V. (1939). On the Estimation of the Discrepancy Between Empirical Curves of Distribution for Two Independent Samples. *Bul. Math. de l'Univ. de Moscou*, 2, 3–14.

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

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.

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