Validation methods for population models of gene expression dynamics

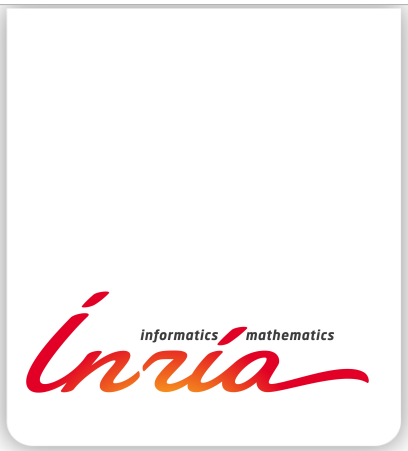
Andres M. Gonzalez-Vargas, Eugenio Cinquemani, Giancarlo Ferrari-Trecate

Validation methods for population models of gene expression dynamics

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

RESEARCH REPORT
N° 8938
July 2016
Project-Team IBIS

ISSN 0249-6399 ISRN INRIA/RR--8938--FR+ENG



Validation methods for population models of gene expression dynamics

Andrés M. González-Vargas*, Eugenio Cinquemani†, Giancarlo Ferrari-Trecate‡

Project-Team IBIS

Research Report n° 8938 — July 2016 — 17 pages

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 report 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.

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

* 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)

RESEARCH CENTRE
GRENOBLE – RHÔNE-ALPES

Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex

Méthodes de validation de modèles de population pour les dynamiques d'expression génique

Résumé : La diffusion des techniques expérimentales pour la mesure de l'expression génique au cours du temps à niveau des cellules individuelles a ouvert la voie à l'étude par modèles de la variabilité intra- et extracellulaire de l'expression génique. Plusieurs approches à l'inférence de modèles de variabilité en populations cellulaires isogéniques ont étés développés et appliques à des contextes réels. Toutefois, moins d'efforts ont étés dédies au développement d'approches systématique a la validation de ces modèles de population, et la qualité des modèles obtenus est souvent évaluée par des critères semi-empiriques. Dans ce rapport on étudie le problème de la validation de modèles des dynamiques des réseaux géniques pour populations cellulaires. On propose des outils statistiques pour la validation et la comparaison qualitative et quantitative de modèles, et on discute leur application et interprétation sur la base d'un problème biologique réel.

Mots-clés : Méthodes statistiques, Biologie des Systèmes, Modélisation stochastique, Modélisation à effets mixtes, Expression génique

1 Introduction

Variability has been recognized to be a crucial aspect of gene expression and regulation [20]. Modern experimental techniques for the monitoring of gene expression at the individual cell level provide both qualitative evidence and quantitative data that can be exploited to describe and analyze gene expression variability from a mathematical standpoint [8, 18]. 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 [17, 25, 18, 15]. 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. Indeed, model assessment is mostly performed on empirical bases, such as qualitative response shape [17, 25], overexpression or knock-out experiments [2], and so on, and it is often limited to considerations pertaining quality of fit, 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 [11]. 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 [13] and Chemical Master Equation [7] 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 $x = [x_1, \ldots, x_n]^T$ 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.

2.1 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 [12, 9]. 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. Population-average measurements y at time t are then described as a function of $x(t)$, i.e.

$$y(t) = f\big(t, u(\cdot), x_0, \psi\big) + \text{measurement noise}$$

where f is determined by the solution of the above ODE for given parameters ψ and initial conditions x_0, under $u(\cdot)$. In the context of population modelling, where single-cell profiles are generally different from one another, the interpretation of this modelling approach needs to be reconsidered. The solution of (1) shall now represent "mean-cell" dynamics, an oversimplification of the ensemble of single-cell responses. If instead the intercellular variability of the responses gives rise to a discrepancy (error) between this mean-cell response and the observations y_i pertaining the ith of N cells, we have

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

2.1.1 Inference.

Inference of a mean cell model can be addressed by a Maximum Likelihood (ML) approach. 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. In particular, an expression of the form

$$h\big(f, \epsilon\big) = \epsilon_a + \epsilon_b f, \tag{3}$$

with $\epsilon = (\epsilon_a, \epsilon_b)$, accounts for a basal error intensity (ϵ_a) plus a contribution proportional to the state ($\epsilon_b f$) and covers most cases of interest. Denoting $\theta = (\psi, \epsilon)$ the set of unknown

parameters (possibly including x_0 as well), the ML estimate of θ may be computed as $\hat{\theta} = \arg\min_\theta -\log \mathcal{L}(\theta|\mathcal{Y})$, where (neglecting unnecessary constants)

$$-\log \mathcal{L}(\theta|\mathcal{Y}) = \sum_{i=1}^{N}\sum_{j=1}^{\mathcal{T}_i}\left\{\frac{1}{2}\left(\frac{y_{i,j} - f\big(t_j, u(\cdot), x_0, \psi\big)}{h\big(f\big(t_j, u(\cdot), x_0, \psi\big), \epsilon\big)}\right)^2 + \log h\big(f\big(t_j, u(\cdot), x_0, \psi\big), \epsilon\big)\right\}.$$

2.2 Mixed-Effects (ME) modelling.

Mean-cell models capture response variability across cells purely in terms of residual modelling error. This is due to the fact that the parametrization of the system dynamics is unique for the whole population. An alternative approach is to assume that (1) models the individual cell, but different cells may be characterized by different values of ψ. In biological terms, this is a basic way to capture what is known as extrinsic noise, i.e. unmodelled sources of variability that alter the kinetics of gene expression, be they environmental or physiological, resulting into similar but non-identical cell responses. If ψ_i denotes the parameters of the ith cell, one then assumes that

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

where $f\big(t, u(\cdot), x_0, \psi_i\big)$ is the solution of (1) with $\psi = \psi_i$, and the error accounts for single-cell model inaccuracy (and measurement noise). Here x is then 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. It is still assumed that u is common across the population, reflecting typical experiments where a given stimulus is applied to a whole population at once. Variability across cells is thus accounted for by the different values of the ψ_i, however, in some analogy with mean-cell models, ensemble population properties of these parameters should still be specified. 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,

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

In particular, it is often assumed that $\psi_i = \mu + b_i$. 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. Individual cell parameter vectors ψ_i are thus mixed effects, containing both fixed and random components. Especially for the case $\psi_i = \mu + b_i$, one may think of μ as the representation of the "mean-cell". Note however, that such cell may well not exist, i.e. no observed profile corresponding to $\psi_i = \mu$.

2.2.1 Inference.

Both μ and the random-effects covariance matrix Ω are population properties, and are generally unknown a priori as are the ψ_i. Inference of mixed-effects models from individual data has the primary aim of reconstructing these population properties from the whole data set $\mathcal{Y}$ of all measurements from all individuals. This is in agreement with the fact that population parameters constitute the information that can be carried over to new cells or experiments, whereas individual cell parameters have their validity limited to the lifespan of a cell. Consider again a generic measurement model of the form (2), where ϵ is fixed across individuals and ψ is replaced by ψ_i.

In a naive, so called *two-stage* approach, one may think of finding, for every i, individual-cell estimates $\hat{\theta}_i = (\hat{\psi}_i, \hat{\epsilon}_i)$ of $\theta_i = (\psi_i, \epsilon)$ from the corresponding profile $\mathcal{Y}_i$ by minimizing the negative log-likelihood

$$-\log \mathcal{L}(\theta_i|\mathcal{Y}_i) = \sum_{j=1}^{\mathcal{T}_i} \left\{ \frac{1}{2} \left(\frac{y_{i,j} - f\big(t_j, u(\cdot), x_0, \psi_i\big)}{h\big(f\big(t_j, u(\cdot), x_0, \psi_i\big), \epsilon\big)} \right)^2 + \log h\big(f\big(t_j, u(\cdot), x_0, \psi_i\big), \epsilon\big) \right\},$$

and then computing estimates for μ and Ω as the empirical mean and covariance of the estimates $\hat{\psi}_1, \ldots, \hat{\psi}_N$. This is an unsatisfactory approach, though, since the information provided by the $\mathcal{Y}_{i'}$, with $i' \neq i$, on ψ_i via the common prior fixed by μ and Ω is not exploited in the estimation of ψ_i. In addition, it is unclear how to relate the various estimates ϵ_i with the population parameter vector ϵ. A better approach, resulting in statistically more accurate estimates of the population parameters $\Theta = (\mu, \Omega, \epsilon)$, is the Population Likelihood Maximization (PLM) approach. The idea is to leverage all data $\mathcal{Y}$ at once by maximizing with respect to Θ the marginal likelihood

$$\mathcal{L}(\Theta|\mathcal{Y}) = p(\mathcal{Y}|\Theta) = \prod_{i=1}^{N} \int d\psi_i p(\mathcal{Y}_i|\psi_i) p(\psi_i|\Theta)$$

(or equivalently minimizing the negative of its logarithm) where factorization occurs thanks to the mutual independence of the b_i and of the η_i. Note that, by this approach, a single estimate is obtained for all population parameters, including ϵ. From the resulting estimates $\hat{\mu}$ and $\hat{\Omega}$, if of interest, one may then derive single-cell parameter empirical Bayes estimates via e.g. Maximum-A-Posteriori (MAP),

$$\hat{\psi}_i = \arg\max p(\psi_i|\hat{\mu}, \hat{\Omega}), \quad i = 1, \ldots, N.$$

In practice, while all integrands can be written explicitly, no closed form expression exists in general for $\mathcal{L}(\Theta|\mathcal{Y})$. Numerical methods for approximate optimization of $\mathcal{L}(\Theta|\mathcal{Y})$ have been proposed (notably NONMEM [1] and the randomized method SAEM [6, 1]) and are contained in dedicated software packages such as Monolix [14], which also provide computation of the $\hat{\psi}_i$ via MAP and other approaches.

2.3 Chemical Master Equation (CME) modelling.

Models above rely on deterministic dynamics for single cells. Therefore, once the individual parameters are fixed, the future evolution of the system becomes fully predictable. This is inadequate when randomness inherent in the biochemical processes of gene expression and regulation is prominent, or simply the main object of study. At the single-cell level, the intrinsic noise in gene expression is captured by modelling the process as a (stochastic) Markov Chain. This approach is more commonly referred to as CME modelling. A CME model is obtained by letting x be a count of molecules of the different species, and interpreting $v(x, \psi)$ as reaction propensities (or intensities), that is, the infinitesimal probabilities that the different reactions may occur in an infinitesimal period of time. If ν_r and v_r are the rth column of ν and the rth row of v, in the same order,

$$\text{Prob}(x(t + dt) - x(t) = \nu_r|\psi) = v_r\big(x(t), \psi\big)dt + o(dt).$$

In turn, rates $v(x, t)$ are again determined by mass-action laws, and ψ are the corresponding kinetic constants [12]. Together with the assumption that the probability of simultaneous events

is of higher infinitesimal order, it follows that $x(t)$ obeys the laws of a Markov chain, and for all possible values z of $x(t)$, the probabilities $p^\psi(z,t) = \mathrm{Prob}\big(x(t) = z\big|\psi\big)$ evolve over time in accordance with the CME

$$\dot{p}^\psi(z,t) = \sum_{r=1}^{m} v_r(z-\nu_r,\psi)p^\psi(z-\nu_r,t) - v_r(z,\psi)p^\psi(z,t), \tag{4}$$

for a given initial probability distribution $p_0(\cdot) = p^\psi(\cdot,t_0)$ at some time t_0. The simplicity of this equation is deceptive: For an infinite space of possible values z, this results in an infinite-dimensional system of coupled linear ODEs, which is analytically intractable except for very few special cases or under nontrivial approximations. Note also that rates may themselves depend on a control input u.

2.3.1 Inference.

In the current literature, CME models are mostly inferred from population snapshot data, that is, from empirical statistics of $z(t)$ computed from independent cell samples at different time points t [17, 25]. 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 [24], but we will not address them here. In the present case, measurements $\tilde{y}(t_j)$ at a sequence of time points $\mathcal{T} = \{t_j : \ j = 1,\dots,T\}$ can be seen as a noisy readout of $p^\psi(\cdot,t_j)$, and the task is to estimate ψ from $\tilde{\mathcal{Y}} = \{\tilde{y}_j : \ t_j \in \mathcal{T}\}$ ("$\tilde{}$" is used here to distinguish measurements of statistics of x from measurements of x itself). For simple enough systems, one approach is to fit approximate solutions of (4)[17, 7], to the sequence of empirical probabilities $\tilde{\mathcal{Y}}$, in the sense of some convenient distance between probability distributions. If the space explored by $x(t)$ with nonzero probability is large, however, this approach is hardly viable. A competing approach of more general applicability is Moment Matching (MM). The idea is to derive from (4) (approximate) dynamical equations for the state moments up to some finite order (typically mean, variance and covariances), and to fit the model-predicted moments to their experimental counterpart. Let $M^\psi(t)$ be the vector containing the moments of $x(t)$ up to order L. It can be shown [21] that

$$\dot{M}^\psi(t) = A(\psi)M^\psi(t) + B(\psi)\bar{M}^\psi(t) \tag{5}$$

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. Matrix B is nonzero except for very few special cases, whence the equation is "open", in the sense that solution depends on the unknown and unmodelled moments $\bar{M}^\psi(t)$. However, several so-called moment closure methods have been proposed to approximate (5) with the "closed" system of equations [21]

$$\dot{\tilde{M}}^\psi(t) = A(\psi)\tilde{M}^\psi(t) + \phi\big(\tilde{M}^\psi(t)\big), \tag{6}$$

where the various methods differ in the definition of ϕ, with an accuracy that is problem-dependent [21, 25]. Let us now model measurements as

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

with usual assumptions on η, and vector c accounting for partial observation of $\tilde{M}^\psi$. In particular, if $L = 2$, then (6) involves mean, variance and covariance terms, whereas only mean and variance

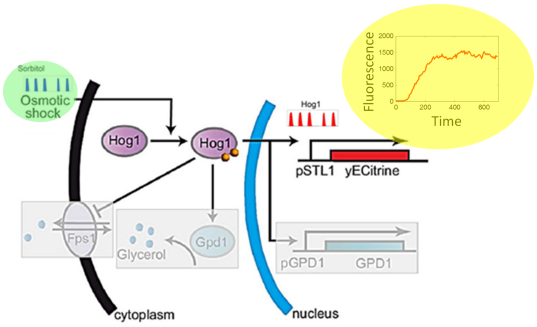


Figure 1: Hyperosmotic gene expression in yeast. Hyperosmotic stress triggers phosphorylation and nuclear import of the protein Hog1, which thereupon activates osmo-stress responsive genes. In [23, 15], a fluorescent reporter gene sequence (yECitrine) is engineered in the cells under the control of osmosensitive promoter pSTL1, which results in the synthesis of fluorescent reporter molecules upon cell sensing of osmotic shocks. 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 the experimental setup in [15, 23], which we take as a reference here, thus Fps1 and GPD1 mechanisms (shaded in light gray) will not be considered.

for a single species are provided by most common experimental setups, such as the one illustrated in this paper. Inference of ψ then becomes the problem of fitting $\tilde{M}^{\psi}$, the solution of (6), to the sequence $\tilde{\mathcal{Y}}$. Assuming (6) exact, an a priori characterization of $h\big(\tilde{M}^{\psi}(t_j), \epsilon\big)\eta(t_j)$ may be given based on the number of cells from which empirical moments are computed from the data [25], and an ML approach can be followed. Otherwise, modelling errors can be implicitly accounted for by including the estimation of ϵ in the identification process. A method of this type based on the Kullback-Leibler (KL) distance has been proposed in [10] for h in the form (3). We will use this method in the sequel without further notification. When discussing validation methods involving single-cell profiles, we will refer back to the single-cell measurements $y_{i,j}$ from which empirical moments $\tilde{y}_j$ are computed. Application of moment-based inference to real biological case studies is reported e.g. in [25, 10].

2.4 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 [15]. 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 [23]). 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 [10], the system is then represented by the following

reactions:

$$\emptyset \underset{k_6}{\overset{k_5 u_h}{\rightleftharpoons}} \text{mRNA} \tag{7}$$

$$\text{mRNA} \xrightarrow{k_7} \text{mRNA} + \text{Protein}^{\text{off}} \tag{8}$$

$$\text{Protein}^{\text{off}} \xrightarrow{k_9} \text{Protein}^{\text{on}} \tag{9}$$

$$\text{Protein}^{\text{off}} \xrightarrow{k_8} \emptyset \tag{10}$$

$$\text{Protein}^{\text{on}} \xrightarrow{k_8} \emptyset \tag{11}$$

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

$$\dot{u}_h(t) = k_h u_c(t) - \gamma_h u_h(t). \tag{12}$$

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 [23]. 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

$$\dot{x}_1(t) = k_5 u_h(t) - k_6 x_1(t), \tag{13}$$

$$\dot{x}_2(t) = k_7 x_1(t) - (k_8 + k_9) x_2(t), \tag{14}$$

$$\dot{x}_3(t) = k_9 x_2(t) - k_8 x_3(t). \tag{15}$$

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

$$y_i(t) = x_3(t) + \big(\epsilon_a + \epsilon_b x_3(t)\big)\eta_i(t). \tag{16}$$

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

3 Validation methods for cell population models

In this section we present various validation criteria that can be used for assessing the quality of cell population models. Moreover, their application will be discussed using the biological example in Section 2.4. To this purpose, we simulate 100 cells using an ME model based on (12)–(16). The osmotic stress profile is shown in Figure 2 (bottom), and it is common to all cells. Outputs are single-cell profiles $y_i(t)$ (Figure 2, top).

Parameters $(k_5, k_6, k_7, k_8, k_9)$ are sampled from a multivariate lognormal distribution, whose mean and covariance matrix, in log-scale, are:

$$\boldsymbol{\mu} = \begin{bmatrix} 3.40 & -1.22 & -0.05 & -5.52 & -4.04 \end{bmatrix}, \quad \boldsymbol{\Omega} = 0.1\mathbf{I}_5.$$

In particular, the values in μ have been adapted from those available in the literature to the particular system described in [10, 15].

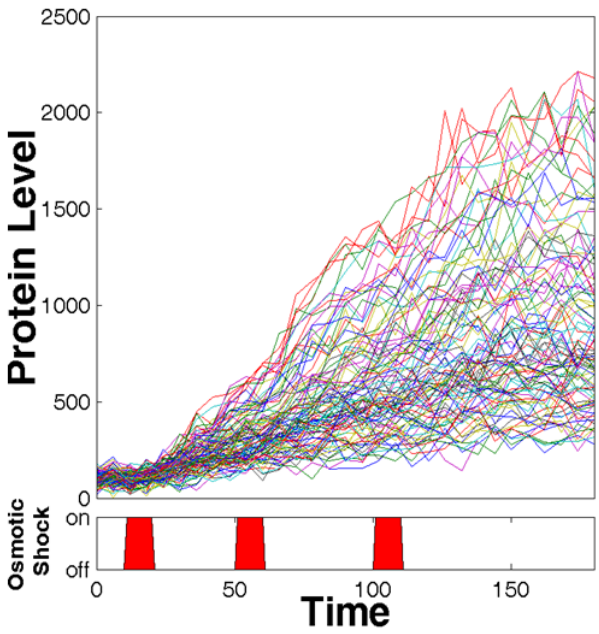


Figure 2: Reference data. Red areas in the bottom plot show the time intervals at which a osmotic shock is exerted. The single-cell (folded) protein levels are shown in the upper plot.

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 3.2 we will describe how the validation criteria can be jointly used for assessing whether a model is acceptable or not.

3.1 Validation criteria for models of gene expression variability

Quality assessment of inferred population models is often non-trivial and it can be easily overlooked if relying only on simple procedures such as visual checks or comparisons of mean and variance of prediction errors. Below, we discuss advanced validation tests capturing the accuracy of population models according to different criteria. In particular, several approaches have been taken from the literature on ME models [19, 4]. While common in classic application fields of ME model (e.g. pharmacometrics), to the authors' knowledge their use in cell-population models is substantially new.

3.1.1 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) captures discrepancies between model predictions and the actual observed values. It 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

$$\mathrm{NRMSE}(\lambda, \hat{\lambda}) = \frac{\sqrt{\frac{1}{T}\sum_{j=1}^{T}(\lambda_j - \hat{\lambda}_j)^2}}{\lambda_{max} - \lambda_{min}} \tag{17}$$

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 (17) denote the set of observed and predicted values, respectively.

3.1.2 Population plots.

A simple way to compare the predicted and observed cell populations is to observe how much the mean and standard deviation of both datasets overlap during the whole experiment duration. To this effect, we compute at every time instant j the empirical mean and standard deviation

$$\hat{m}_{\mathcal{Y},j} = \tfrac{1}{\mathrm{N}_j} \textstyle\sum_{i \in \mathcal{N}_j} \mathcal{Y}_{ij} \tag{18}$$

$$\hat{\sigma}_{\mathcal{Y},j} = \sqrt{\tfrac{1}{\mathrm{N}_j - 1} \textstyle\sum_{i \in \mathcal{N}_j} (\mathcal{Y}_{ij} - \hat{m}_{\mathcal{Y},j})^2} \tag{19}$$

A dataset of simulated cells is then created using the identified model. Formulae (18) and (19) are then used to calculate the predicted mean ($m_{y,j}$) and standard deviation ($\sigma_{y,j}$) of the population by replacing observed data $\mathcal{Y}$ with simulated data y. 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*. Figure 3 shows standard plots for the models of interest.

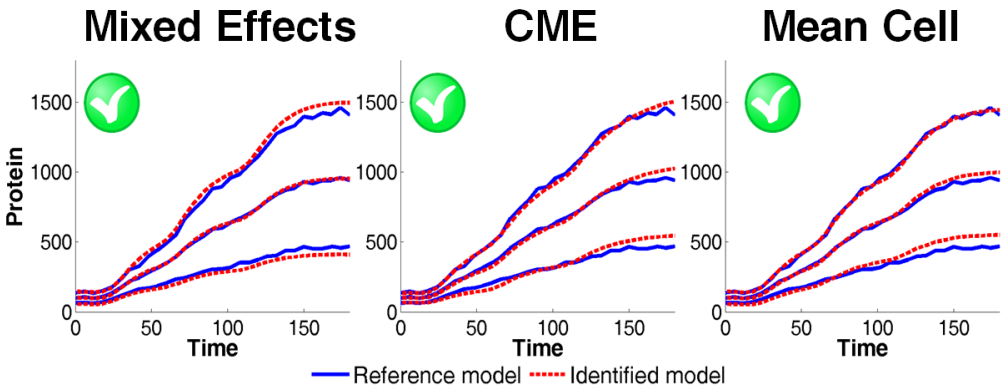


Figure 3: Standard plots: mean +/- standard deviation of reference (blue) and predicted (red) datasets.

The standard plot provides information about the location and dispersion of the population, but its main drawbacks are that it implicitly assumes normality in the distribution of the output across the population (it only accounts for mean and variance). In addition, it does not take into account single-cell fits.

3.1.3 Visual Predictive Check (VPC).

VPC is a popular method for evaluating nonlinear ME models in population pharmacometrics [4, 13]. 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. Indeed, it is important to

remember that, due to the technical complexity of experiments, the number of observed cells is often small (<1000), which affects the accuracy of empirical moments or estimated quantiles. 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. Several extensions have been proposed, but in this paper we will use only the simplest version of the VPC. Figure 4 shows the classic VPC for the models of interest.

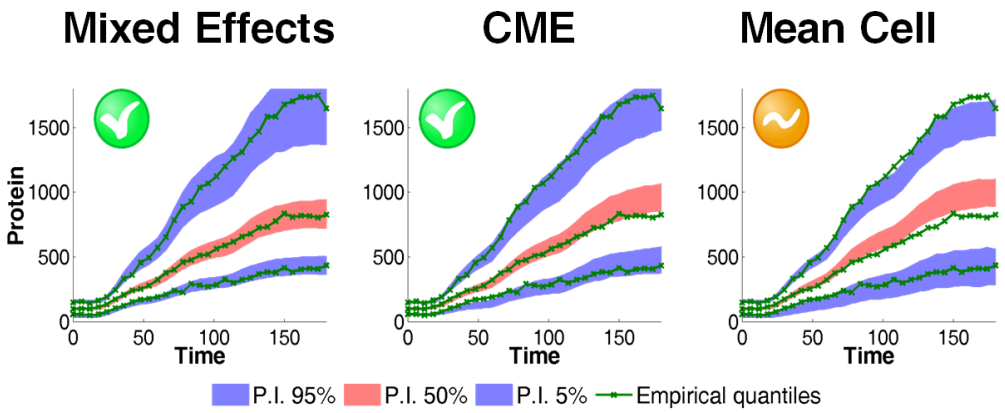


Figure 4: VPC: shaded areas denote 99% confidence intervals on the calculated quantiles for the predicted dataset. The selected quantiles are 0.025 (blue), 0.5 (red) and 0.0975 (blue) which comprise 95% of the population. The green lines show the same quantiles for the reference dataset. A large deviation of the reference quantiles from the predicted quantiles' area suggests misspecification in the model.

3.1.4 Kolmogorov-Smirnov test.

The Kolmogorov-Smirnov Two-Sample Test (KS2) is used to assess, without assumptions of the underlying probability distributions, the similarity between two sample distributions [22].

The KS test is based on the Empirical Cumulative Distribution Function (ECDF). Let X be a scalar random variable and $X_1, \dots, X_N$ be a sample of X. The ECDF is defined as:

$$F_N(x) = \frac{1}{N} \sum_{i=1}^{N} \Upsilon_x(X_i) \tag{20}$$

where $x \in \mathbb{R}$ is typically in the interval $[\min(X_i), \max(X_i)]$, and $\Upsilon_x(X_i)$ is 1 if $X_i \leq x$, and 0 otherwise. 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 ECDFs of the observed and simulated datasets. Then we compute

$$D_{o-p} = \sup_x |F_{1,N}(x) - F_{2,N'}(x)| \tag{21}$$

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

$$D_{N,N'} > c(\alpha) \sqrt{\frac{N+N'}{NN'}}. \tag{22}$$

If we choose a significance level of 95% then $\alpha = 0.05$ and $c(\alpha) = 1.36$ [16]. 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 time interval of the experiment, by computing hK_j at each time instant j and then taking the mean over all time instants:

$$S_{ks} = 1 - \frac{1}{T} \sum_{j=1}^{T} hK_j \tag{23}$$

We can also compute the average p-value of S_{ks}. A higher p-value will indicate that the two distributions are more similar. Figure 5 gives a graphical representation of the KS2 statistic.

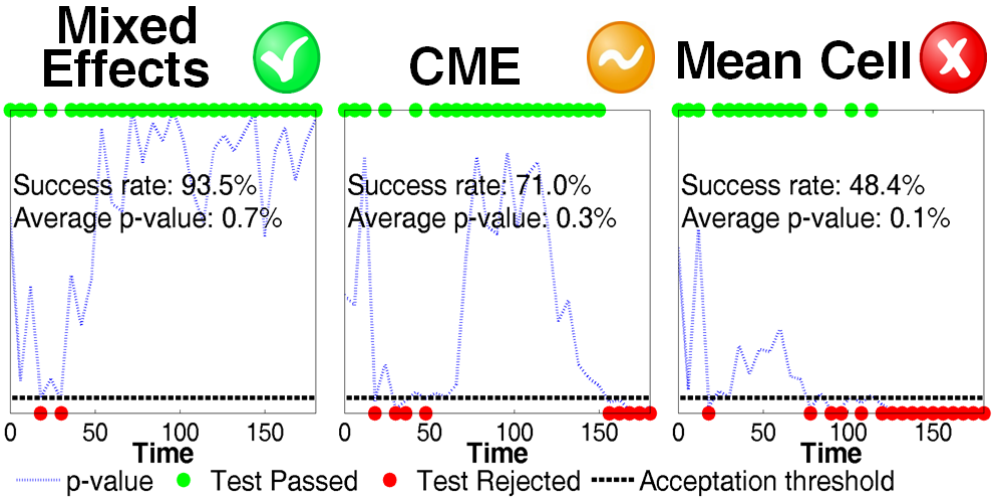


Figure 5: KS2 test. The blue line represents the p-value obtained from the test at each time instant (the higher the better). The 95% threshold p-value (black-dashed line) separates unsuccessful time points (red points, indicating the distributions are statistically different) from successful time points (green).

3.1.5 Prediction distribution errors

The Prediction Distribution Errors (PDE) are proposed in [4] as a metric to evaluate the performance of a ME model, based on Monte Carlo simulations of the population. Normalized PDE (NPDE), a variant of PDE, is widely used in pharmacokinetics, and is implemented in statistical software such as R and Monolix.

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. This requires to derive an approximated variance-covariance matrix for single-cell data. Decorrelation can then be performed using several methods, such as Cholesky decomposition, inverse decorrelation through eigenfunctions, or singular value decomposition [3]. 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 (prediction distribution error) as:

$$\mathrm{PDE}_{ij} = \frac{1}{K} \sum_{k=1}^{K} \delta_{ijk}^*, \tag{24}$$

where $\delta^*_{ijk} = 1$ if $y_{ij}^{\text{sim}(k)*} < \mathcal{Y}^*_{ij}$ 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 can be obtained by using the inverse function of the normal cumulative density function F: $\text{NPDE}_{ij} = F^{-1}(\text{PDE}_{ij})$.
Results of the NPDE can be seen in Figure 6. 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 integrates the result of the combination of three different normality tests: Wilcoxon, Chi-square and Lilliefors [5]. Together, they give 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 Figure 6.

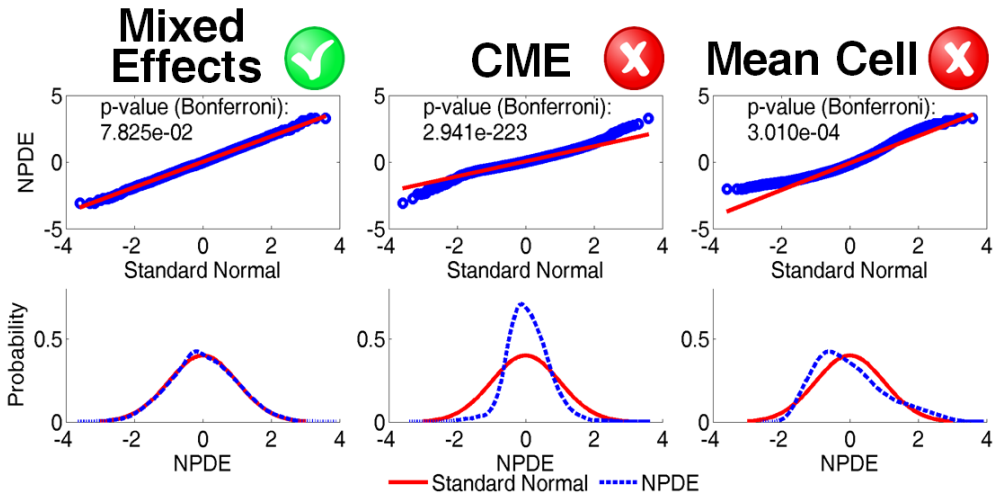


Figure 6: Normalized prediction distribution errors (NPDE). Quantile-Quantile plots (top) compare the NPDE distribution (blue circles) to a normal standard distribution (red line). The Bonferroni-corrected p-value quantifies the closeness of both distributions. The bottom plots show the same comparisons, but from the perspective of probability density functions.

3.1.6 A posteriori best fits (APBFs).

Using the simulated datasets introduced for discussing PDEs, we can compute, for observed cell i,

$$\text{APBF}_i = \arg\min_k(\text{NRMSE}(y_i^{sim(k)}, \mathcal{Y}_i)) \tag{25}$$

In other words, APBF_i denotes the index k that minimizes the NRMSE between $y_i^{\text{sim}(k)}$ 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):

$$\text{NRMSE}_{\text{APBF}} = \tfrac{1}{N}\textstyle\sum_{i=1}^{N} \text{NRMSE}(y_i^{\text{sim}(\text{APBF}_i)}, \mathcal{Y}_i) \tag{26}$$

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 $\text{NRMSE}_{\text{APBF}}$, the better the model is able to represent individual cells. Fig. 7 shows APBF plots for the models of interest.

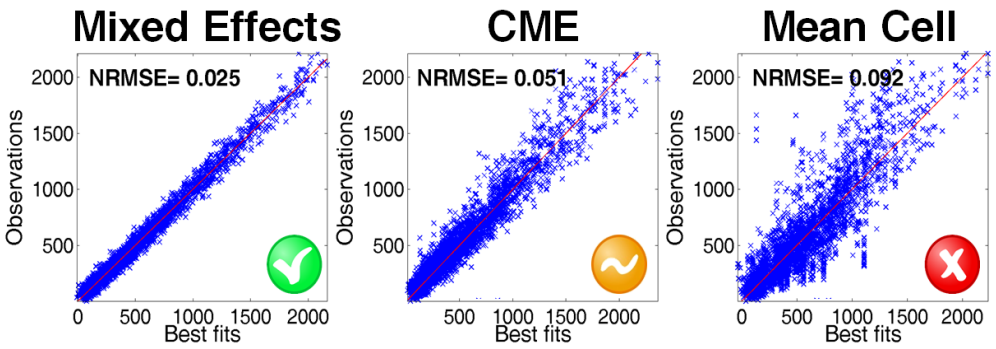


Figure 7: APBF plots. Points $(y_{ij}^{\mathrm{sim}(\mathrm{APBF}_i)}, y_{ij})$ for $i = 1, \ldots, N$, $j = 1, \ldots, T_i$ are represented, i.e. individual best predictions against observed values of the reference data. A lower spread of the points in the anti-diagonal direction indicates better agreement between observations and predictions; this can be quantified by calculating $\mathrm{NRMSE}_{\mathrm{APBF}}$ as in (26).

3.2 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 3.1, 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 Figures 3-7. For an easier visual comparison between models, we display in all figures colored circles indicating good (green), moderate (yellow) and bad (red) results.

The first evaluation (Figure 3) is a comparison of standard plots. Based on this criterion, all models seem to perform equally well. The mean is followed more closely by the ME model and the variance seems to be overestimated in ME, and slightly off in the last part of MC and CME, but it is difficult to provide any strong evidence in favor of a model.

The second test is yet another visual evaluation based on VPC (Figure 4). The green lines representing the empirical quantiles of the reference data fall always inside the limits of the quantiles predicted with the ME and CME models, but tend to fall outside of the MC predicted quantiles. This gives us some preliminar evidence of model misspecification in the MC case.

The third test is a Kolmogorov-Smirnov 2-sample test (Figure 5). Here, we see that the KS2 success rate for the ME model is 93.5%, while those of CME and MC are 71% and 48.4% respectively. This gives us strong evidence to discard the MC model, and suggests that the CME model is also not valid with high significance.

The previous tests all evaluated the capacity of the identified model to reproduce a *population of cells* that behaves similarly to the reference dataset. However, the true model should also be able to reproduce an individual cell with good quality. This aspect is covered by the next two tests.

The fourth test (Figure 7) compares each of the reference cells to their correspondent best-fitting cell from the predicted dataset. A visual analysis of this test tells us that if the predicted model is able to fit sufficiently well the individual cells, all the blue crosses in Fig. 7 should be very close to the diagonal. A larger dispersion in the anti-diagonal direction means that residuals will be larger. The best model should show little dispersion and the NRMSE should be as low as possible. The NRMSE indicates that ME is better than the two competing models.

The last test is the NPDE, and, in some sense, it evaluates simultaneously the individual and population performance of the model. The Q-Q and PDF plots in Figure 6 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. If the evidence is not conclusive, numerical indicators such as KS2 can help to assess the performance of the identified models at the population level. However, performance at the single-cell level is equally important and it must be confirmed by analyzing the best-fit residuals.

4 Conclusions

In this paper, we have compared and contrasted methods for validating models of cell populations. Overall, our analysis shows that tests based on the capability of reproducing only population-level behaviors might be insufficient for model discrimination. To this purpose, it is beneficial to consider also validation methods based on the comparison of single-cell data. 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, we expect they will be further developed in the future, so as to incorporate genuine biological aspects in their formulation.

References

[1] Robert J. Bauer, Serge Guzy, and Chee Ng. A survey of population analysis methods and software for complex pharmacokinetic and pharmacodynamic models with examples. *AAPS JOURNAL*, 9(1), 2007.

[2] Irene Cantone, Lucia Marucci, Francesco Iorio, Maria Aurelia Ricci, Vincenzo Belcastro, Mukesh Bansal, Stefania Santini, Mario di Bernardo, Diego di Bernardo, and Maria Pia Cosma. A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. *Cell*, 137(1):172 – 181, 2009.

[3] E. Comets, K. Brendel, and F. Mentrè. Computing normalised prediction distribution errors to evaluate nonlinear mixed-effect models: The NPDE add-on package for R. *Computer methods and programs in biomedicine*, 90(2):154–166, 2008.

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

[5] William Jay Conover and WJ Conover. Practical nonparametric statistics. 1980.

[6] Bernard Delyon, Marc Lavielle, and Eric Moulines. Convergence of a Stochastic Approximation Version of the EM Algorithm. *The Annals of Statistics*, 27(1):94–128, 1999.

[7] Hana El Samad, Mustafa Khammash, Linda Petzold, and Dan Gillespie. Stochastic modelling of gene regulatory networks. *International Journal of Robust and Nonlinear Control*, 15(15):691–711, 2005.

[8] Michael B. Elowitz, Arnold J. Levine, Eric D. Siggia, and Peter S. Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, 2002.

[9] Daniel Gillespie. The chemical Langevin equation. *Journal of Chemical Physics*, 2000.

[10] Andres M Gonzalez, Jannis Uhlendorf, Joe Schaul, Eugenio Cinquemani, Gregory Batt, and Giancarlo Ferrari-Trecate. Identification of biological models from single-cell data: a comparison between mixed-effects and moment-based inference. In *Proceedings of the 12th ECC*, pages 3652–3657, April 2013.

[11] Ryan N Gutenkunst, Joshua J Waterfall, Fergal P Casey, Kevin S Brown, Christopher R Myers, and James P Sethna. Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput Biol*, 3(10):1–8, 10 2007.

[12] R. Heinrich and S. Schuster. *The regulation of cellular systems*. Springer, 1996.

[13] M. Lavielle. *Mixed-Effects models for the population approach*. CRC press, 2015.

[14] Lixoft. *Monolix User Manual Version 4.3.2*. Lixoft, November 2014.

[15] Artemis Llamosi, Andres M. Gonzalez-Vargas, Cristian Versari, Eugenio Cinquemani, Giancarlo Ferrari-Trecate, Pascal Hersen, and Gregory Batt. 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, 02 2016.

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

[17] Brian Munsky, Brooke Trinh, and Mustafa Khammash. Listening to the noise: random fluctuations reveal gene network parameters. *Molecular Systems Biology*, 5, October 2009.

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

[19] José C. Pinheiro and Douglas M. Bates. *Mixed-Effects Models in S and S-PLUS*. Springer Verlag, New York, first edition, 2000.

[20] C. V. Rao, D. M. Wolf, and A. P. Arkin. Control, exploitation and tolerance of intracellular noise. *Nature*, (420):231–237, 2002.

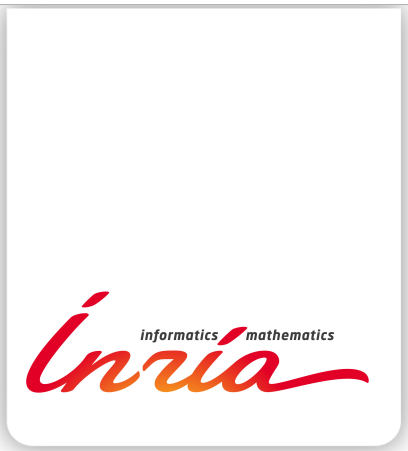
[21] David Schnoerr, Guido Sanguinetti, and Ramon Grima. Comparison of different moment-closure approximations for stochastic chemical kinetics. *The Journal of Chemical Physics*, 143(18), 2015.

[22] N. V. Smirnov. 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, 1939.

[23] Jannis Uhlendorf, Agnès Miermont, Thierry Delaveau, Gilles Charvin, François Fages, Samuel Bottani, Gregory Batt, and Pascal Hersen. Long-term model predictive control of gene expression at the population and single-cell levels. *PNAS*, 109(35):14271–14276, August 2012.

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

[25] Christoph Zechner, Jakob Ruess, Peter Krenn, Serge Pelet, Matthias Peter, John Lygeros, and Heinz Koeppl. Moment-based inference predicts bimodality in transient gene expression. *PNAS*, 109(21):8340–8345, May 2012.



RESEARCH CENTRE
GRENOBLE – RHÔNE-ALPES

Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399