INTERFACE

rsif.royalsocietypublishing.org

Review

Cite this article: Waldherr S. 2018 Estimation methods for heterogeneous cell population models in systems biology. *J. R. Soc. Interface* **15**: 20180530.
http://dx.doi.org/10.1098/rsif.2018.0530

Received: 12 July 2018
Accepted: 4 October 2018

Subject Category:
Life Sciences–Engineering interface

Subject Areas:
computational biology, systems biology

Keywords:
parameter identification, state estimation, population balance equations, structured population models, density functions, cell ensemble models

Author for correspondence:
Steffen Waldherr
e-mail: steffen.waldherr@kuleuven.be

THE ROYAL SOCIETY
PUBLISHING

Estimation methods for heterogeneous cell population models in systems biology

Steffen Waldherr

Department of Chemical Engineering, KU Leuven, Leuven, Belgium

SW, 0000-0002-0936-579X

Heterogeneity among individual cells is a characteristic and relevant feature of living systems. A range of experimental techniques to investigate this heterogeneity is available, and multiple modelling frameworks have been developed to describe and simulate the dynamics of heterogeneous populations. Measurement data are used to adjust computational models, which results in parameter and state estimation problems. Methods to solve these estimation problems need to take the specific properties of data and models into account. The aim of this review is to give an overview on the state of the art in estimation methods for heterogeneous cell population data and models. The focus is on models based on the population balance equation, but stochastic and individual-based models are also discussed. It starts with a brief discussion of common experimental approaches and types of measurement data that can be obtained in this context. The second part describes computational modelling frameworks for heterogeneous populations and the types of estimation problems occurring for these models. The third part starts with a discussion of observability and identifiability properties, after which the computational methods to solve the various estimation problems are described.

1. Introduction

Organisms and microbial systems are defined by the combined actions of cells from large populations within a tissue or microbial colony. Even within a genetically homogeneous population, individual cells often show a considerable amount of phenotypic heterogeneity. The review [1] discusses several cases in which the cell heterogeneity needs to be known to fully understand the biological mechanisms at play. Examples are cell differentiation or other fate decisions, and subpopulations of cells with a different phenotype within a genetically identical population. Studies on the effects of this heterogeneity generally revealed that organisms have evolved to cope with or even profit from cellular variability, for example, by improving fitness in changing environments [2] or allowing persistence under environmental stress. On the other hand, cellular heterogeneity can be undesired from a bioprocess or medical perspective, as it may for example reduce the productivity of biotechnological processes [3,4] or lead to the emergence of tumour resistance during cancer therapy [5]. The relevance of cellular heterogeneity for both basic understanding of living systems and applications has motivated the development of experimental techniques that can elucidate heterogeneity by high-throughput measurements on single cells, as well as the construction of computational models to simulate the emergence, dynamics and effects of cellular variability.

Regarding the experimental techniques, understanding cellular heterogeneity requires individual measurements of a sufficiently large number of single cells. Single-cell analysis techniques have been under development for several decades and have made significant progress in the last decade. A comprehensive overview of current single-cell analysis techniques is given in [6]. Two basic approaches are of particular relevance for estimation problems with dynamic computational models: flow cytometry, which gives rise to non-time resolved, the so-called population snapshot data [7], and time lapse microscopy using either cell tracking with image analysis [8] or microfluidics [9] to keep track of individual cells over time.

This review discusses methods for estimation problems arising from computational models of heterogeneous cell populations. In a stochastic setting, cell heterogeneity can be described by looking at a probability distribution over the cellular variables, for example, through the chemical master equation [10,11]. A detailed overview on single-cell stochastic models is provided in [12].

In the so-called 'population balance equations' (PBEs), a density function over the heterogeneous cellular variables is being formulated, and the change of this density function over time describes the population dynamics [13]. Mathematically, PBEs are partial differential (or integro-differential) equations, with time and the heterogeneous cellular variables as independent variables. These models contain distinct elements for several processes which together shape the population heterogeneity, and are thus useful, for example, to distinguish the effects of different parts of the system including intracellular dynamics, cell division and death rates or cell partition at division, on the overall population state and dynamics. A detailed exposition of PBEs is given in [14], and [15,16] review more applications of these models to biological systems.

Another option to model cell population behaviour are individual-based or cell ensemble models [15,17,18]. These models simulate a large number of individual cells which are representative for the full population. Population characteristics can then be inferred from gathering the individual cells' properties. While individual cells are sometimes modelled stochastically in ensemble models, the statistical properties of the overall population do still deterministically depend on the single-cell dynamics.

Estimation problems that are being discussed here mainly involve the estimation of state variables and parameters in computational models from appropriate experimental data. State and parameter estimations are long-standing issues in computational modelling and are of particular importance for biological models, where modelling from first principles alone is mostly not possible. For classical models based on ordinary differential equations (ODEs), state and parameter estimation methods are very well established. The tutorial [19] gives an overview of state estimation techniques for this type of models, with a focus on chemical and biochemical processes, which are mostly classic but still the most commonly used at present. An application of these state estimation techniques to parameter estimation for ODE models has been discussed by Lillacci & Khammash [20]. This also covers the problem of identifiability analysis, i.e. the question whether the available measurements are sufficient to uniquely determine parameter values. For more specific model types of biochemical reaction networks within the ODE framework, for example, linlog models [21] or mass-action reaction networks [22], even analytical methods for identifiability analysis are available.

In contrast to the situation with ODE models, state and parameter estimation for cell population models has, with the exception of a few early publications, only found more widespread research attention in the past few years. The purpose of this review is to provide an overview on recent developments in this field, and to elaborate the relationship between experimental measurement techniques with computational cell population models and the associated numerical estimation techniques.

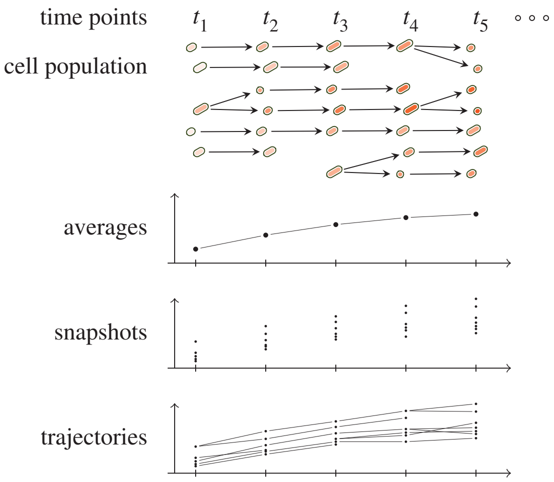

Figure 1. Illustration of three experimentally obtainable types of measurements for cell populations: population averages, population snapshots and cell trajectories. In this example, measurements are taken at five different time points. The panel labelled 'cell population' shows the actual cell population dynamics for this illustration, with colour intensities representing the cellular variable available for measurement. The panels below show which information can be obtained through the different types of measurements. Different cell sizes are included in the cell population, but not in the measurement data. (Online version in colour.)

While spatial organization and effects of mechanical interactions often play an important role for cell populations, and various modelling frameworks exist to deal with them [23], this review does not discuss these aspects explicitly. Owing to the associated experimental challenges, suitable data that would include spatial information for model-based estimation are often not available. Nevertheless, in some cases models and estimation methods can be extended to spatial models in a straightforward way, for example, by including position information with the cellular variables.

This review is structured in three main parts. First, the types of experimental data that can be used in estimation problems for cell population models are introduced. The second part presents model classes that are commonly used for simulating cell populations, together with the associated estimation problems. The third part is an overview on the common estimation methods, and how they apply to the different models and estimation problems discussed in the second part.

2. Experimental data used for estimation with cell population models

Classical quantification techniques in molecular biology such as the various blotting methods or microarrays measure a variable as a population average only, and are thus of limited utility for cell population models. This measurement type is illustrated in the panel labelled 'averages' in figure 1. While such measurements can of course be compared to the population averages that are also readily computed in the population models, they are not rich enough to fully characterize a heterogeneous population state, and thus can only be used as auxiliary, but not as core data for population estimation.

Estimation for cell population models requires measurements from individual cells, where a cellular property of interest is quantified on a single-cell basis for a large number of cells from the population. A widely established experimental method for such data is flow cytometry, where a large number of cells (can be hundred thousands) are streamed individually through a laser beam, and the refraction properties (forward and side scatter) can be correlated with cell size and morphology [24]. Also fluorescence measurements can be integrated, so that for example protein levels can be measured by a fluorescent tag or labelling with a fluorescent antibody [24]. A more recent technology are the so-called image stream systems, where a large number of cells (typically ten thousands) are streamed individually through a camera system, which takes a microscopic image of each cell [25]. With image analysis methods, again cell sizes and protein levels through fluorescent labelling can be determined. In addition to flow cytometry, such a system also permits one to quantify the distribution of proteins between different cellular compartments, such as cytosol and nucleus, which is relevant for many intracellular signal transduction pathways.

Another relevant recent development is the simultaneous measurement of mRNA and protein levels in single cells, either through antibody staining and flow cytometry for a relative quantification [26], or through a combination of proximity ligation assay, polymerase chain reaction and a droplet emulsification at limiting dilution for digital quantification [27].

A key characteristic of cell populations is the population size, and experimental approaches to specifically quantify the cell division dynamics have been developed. A simple approach to obtain population sizes and their changes is just by counting the numbers of cells in a defined sample volume at different time points during an experiment, for example, with a Coulter counter. A more detailed view on the population dynamics is obtained through flow cytometry, after labelling the cells with a persistent label. A persistent cell labelling is a detectable (fluorescent, for example) marker that is brought into cells at the start of an experiment and is persistently incorporated into cellular structures. Examples are carboxyfluorescein succinimidyl ester (CFSE), which is a fluorescent dye that is covalently linked with intracellular molecules, or bromodeoxyuridine (BrdU), which is incorporated into the cells' DNA [28]. The label amount is diluted through cell divisions, and thus the reduction of the label compared to the initial staining is a measure for the generation number of each cell. The fluorescence measurements provide population snapshots for the label amounts per cell, yielding a label histogram or density function as discussed above. Ideally, a heterogeneous population will give rise to multiple peaks in the histogram, from which the population ratios for the different generation numbers can be determined.

Laser refraction and fluorescence measurements (imaging or with a photo-sensor) have the disadvantage that relatively few parameters are simultaneously accessible for measurement. In principle, 10–20 fluorescence channels could be used simultaneously [29]; however, practical hurdles such as incompatibility of immunostaining protocols or antibody availability and specificity often limit this approach to a handful simultaneously measured variables per cell. Using advanced strategies such as cyclic immunofluorescence with fixed cells, fluorescent immunostaining can be extended to around 15 measured variables per cell [30]. Through the more recently developed, but not yet widely used, mass cytometry, dozens of parameters per single cell can be measured more reliably [31]. Still, these numbers are not yet on a par with the different 'omics' approaches for population average measurements, where easily thousands or more simultaneously measured variables are available. With the increased interest in measurements at single-cell resolution, different 'omics'-type measurement techniques have been developed for single cells as well [6,29,32]. For single-cell transcriptomics [33,34], indeed thousands of genes can be quantified simultaneously. Typical cell numbers used in studies over the past years ranged from the multiple tens [34] to few hundreds [35], but more recent microfluidic technology using gel beads allows one to scale that up to multiple thousand cells per experiment [36].

For protein quantification, different techniques such as Western blotting [37] or antibody arrays [38] have been developed for single-cell measurements. While the achievable cell numbers range into the thousands [37], the number of proteins that can be measured simultaneously is about a dozen, and thus still rather limited compared to bulk proteomics.

A common feature of the experimental approaches discussed so far is that cells are sacrificed for the measurement, either for fixation or lysis, or that, even if the cells survive, the perturbation through the measurement is so high that an individual cell cannot be reused in the experiment after the measurement. That means that only a single measurement, i.e. one time point over the course of an experiment, can be obtained from each cell. While measurements on a subpopulation can be taken at different time points during an experiment of course, the correlation between time points from single cells cannot be seen from these experiments. Instead, only independent population samples can be obtained from different time points. For this reason, these types of data have been called 'population snapshots' in past studies [7]. This is illustrated in the accordingly labelled panel of figure 1. From such population snapshots, histograms or density functions can be generated that represent the distribution of the measured variables through the cell population at the time point where the measurement was taken. Alternatively, moments of the underlying distribution can be computed from the individual data points in the snapshot. The laser refraction measurement principle underlying flow cytometry has also been transferred to *in situ* bioreactor probes, which permit one to measure, for example, cell size distributions or viability parameters online at single-cell resolution during a biotechnological process [39,40]. These measurements also provide snapshot data only.

In order to maintain the correlation among different time points for individual cells and to obtain single-cell time courses, time-resolved measurements from the single cells are required. Such measurements are done either by tracking cells through image analysis methods [8,41] or by trapping them in a specific position on a microscale analysis platform [6]. Here, the cell numbers are still low (typically hundreds), but full temporal trajectories of single-cell variables are obtained, as illustrated in the panel labelled 'trajectories' of figure 1. By tracking cellular mother–daughter relations through division events, full cellular lineage trees can be constructed [42,43], which allow one to extend the temporal correlation for single-cell properties even across cell divisions. It should be noted though that temporal tracking of single-cell properties requires a dedicated experimental set-up and cannot be used in all situations. Typically, this can only be realized in laboratory cell culture set-ups, and will not be available to characterize cell populations *in vivo* or online in a biotechnological process.

3. Cell population models and estimation problems

Cell population models are mathematical models which describe the dynamics of a large number of living cells. In contrast to classical population models, where only the population size is described, the models considered here include heterogeneity by allowing individual cells within a potentially large population to take on different states. These models may include both population dynamics through cell division and death, which change the number of cells, as well as cellular dynamics, which change the state of individual cells. The cellular dynamics typically arise from intracellular processes in metabolism, signalling or gene regulation.

The objective of this section is to introduce modelling approaches for heterogeneous cell populations as well as related estimation problems in systems biology. It is not meant to be a comprehensive modelling review, since reviews more specific to the different modelling frameworks are already available in the literature. For each model class, first the modelling approach is described, and then the estimation problems relevant to that model class are discussed.

3.1. General properties of models and estimation problems

For each model, one needs to distinguish between the model's *free variables*, which can be chosen independently of the model's equations, and the model's *dependent variables*, which can be computed from the free variables by solving the model equations. Typical free variables that are relevant for estimation are constant parameters and initial and boundary conditions for the system's configuration. The dependent variables are frequently describing the evolution of the population state after the initial time.

For the definition of the estimation problem, it is obviously necessary to specify which elements of the model are unknown and should be estimated from experimental data. Estimation is only applicable to the model's free variables, as the dependent variables are already determined through the model equations which underly the estimation problem. Note that frequently not all free variables need to be estimated, because some of them might be known from *a priori* knowledge or direct measurements. A problem where the model's initial configuration is to be estimated is called a *state estimation* or *observation* problem, while the estimation of model parameters is called a *parameter estimation* or *identification* problem.

A crucial distinction to be made here is about the mathematical nature of the elements to be estimated. In the context of population models, one can estimate real-valued parameters or variables as in more classical estimation problems, but one can also estimate functions. In the so called *non-parametric estimation*, no specific mathematical expression for the function is assumed, but the goal is to determine the function itself. Typically, the function is numerically discretized on a grid or with appropriate basis functions.

The model is connected to the measurements by means of a *measurement equation*, which describes the measured values as a function of the model's state. Attention needs to be given to the measurement noise affecting the observations. For some estimation methods, a model of the measurement noise is required in the estimation procedure. Likelihood-based methods, for example, use the noise model to determine the likelihood function, and the quality of the estimation results and associated uncertainties may depend on the quality of the noise model.

3.2. Models based on the chemical master equation

3.2.1. Description of the model class

Stochasticity of chemical reactions is a common source of heterogeneity among cells in a population. When considering discrete cellular states (such as molecule numbers), it can be modelled by a (chemical) master equation (CME). This is a differential equation that describes the probabilities for each of the cellular states to occur, and their evolution over time [10]. In a CME model, dependent variables are thus the probabilities for all of these states.

Mathematically, the CME is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}P(t,x) = \sum_{j=1}^{m}(\nu_j(x-v_j,\theta)P(t,x-v_j) - \nu_j(x,\theta)P(t,x)), \tag{3.1}$$

where x is an n-dimensional vector of non-negative integers representing the possible cellular states and $P(t, x)$ their probability to occur. The model includes m transitions for each state, for example, chemical reactions, each occurring with a propensity ν_j that depends on state where it originates from and some parameters θ. The stoichiometric vectors v_j describe the change in state x occurring upon the transition j.

When solving estimation problems, the CME mostly needs to be approximated by computationally efficient simulation methods. A finite-dimensional differential equation can be obtained in two principal ways. One is by truncating the state space using the so-called finite-state projection [44]. It leads to a large scale but linear ODE of the form

$$\dot{P} = A(\theta)P, \tag{3.2}$$

where P is the vector of probabilities for each state that is retained in the truncation, A is a matrix of coefficients for transition propensities, and θ contains parameters such as kinetic constants. The dimension of P in (3.2) typically is still many thousands or more, so for parametric problems such as parameter estimation requiring repeated simulations, it may be appropriate to reduce the model dimension further. A significant reduction of the model dimension can, for example, be achieved by parametric reduced basis methods, which retain all parameters θ in the reduced model. With this approach, reductions from a dimension of about 22 000 to 33 and from about 90 000 to 109 have been achieved in two case studies [45].

The other way to obtain a finite-dimensional differential equation of low to moderate dimension is to derive a differential equation for the evolution of the moments instead of the full probability distribution. This will yield a differential

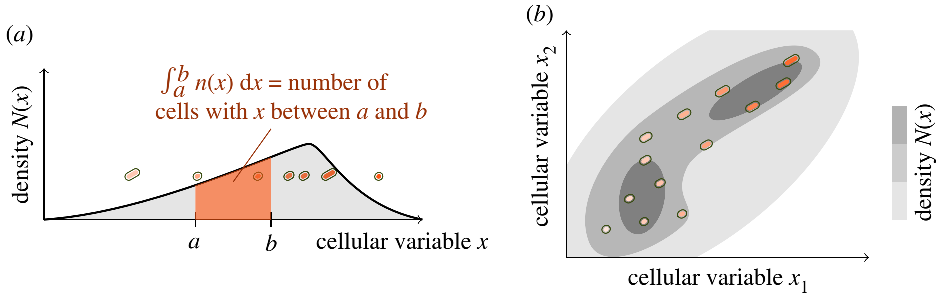

Figure 2. Illustration of the number density function $N(x)$. (a) One-dimensional cellular variable x, represented by colour intensity in the symbolic cells. (b) Two-dimensional cellular variable x, where x_1 corresponds to colour intensity and x_2 to cell size. (Online version in colour.)

equation of the form

$$\frac{\mathrm{d}}{\mathrm{d}t}\mu = F(\mu, \theta), \tag{3.3}$$

where μ is the vector of moments defined by $\mu_i(t) = \sum_x P(t, x)x^i$. If either the number of species or the order of the moments is low, then (3.3) will be of low dimension compared with the number of possible discrete states. If some of the propensities ν_j are nonlinear in x, then (3.3) will not be a closed system, and (approximate) moment closure techniques will have to be used to obtain a finite-dimensional system [46,47].

For CME-based models, using the finite-state projection (3.2) can give a quite accurate solution, but unless molecule numbers are very low, it is only feasible for systems with up to three molecular species. However, using various approximation techniques such as reviewed in [48], generating a sample or computing moments becomes computationally feasible for most systems that can also be solved deterministically based on a kinetic model. Several of these approximations rely on interpreting the states x as a continuous instead of an integer vector. With that approach, the probability mass function $P(t,x)$ is replaced by a probability density function $p(t,x)$, and the dynamics of $p(t,x)$ are given by a Fokker–Planck equation of the form

$$\begin{aligned}\frac{\partial p}{\partial t}(t, x) &= -\mathrm{div}(f(x, \theta)p(t, x))\\ &\quad + \frac{1}{2}\sum_{i,j=1}^{n}\frac{\partial^2}{\partial x_i \partial x_j}(B_{ij}(x, \theta)p(t, x)),\end{aligned} \tag{3.4}$$

where $f(x, \theta) = \sum_{j=1}^{m} v_j \nu_j(x, \theta)$ is a drift term and $B(x, \theta) = \sum_{j=1}^{m} v_j v_j^{\mathrm{T}} \nu_j(x, \theta)$ is an $n \times n$ diffusion matrix [48,49]. We will see in §3.3 that the Fokker–Planck equation (3.4) closely relates to models based on a PBE.

3.2.2. Related estimation problems

A prototypical estimation problem involving CME models assumes having population snapshot data in the form a measurement of the probability distribution $P(t, x)$ at multiple time points, and aims at estimating values for the parameters θ in the model (3.1). Such a direct use of snapshot data is usually restricted to low-dimensional state vectors x, with one or two cellular variables [50]. It is also possible to consider measurements of the moments $\mu(t)$ at multiple time points, from which the parameters θ can be estimated.

In population models based on the CME, one usually distinguishes intrinsic noise generated by the particular chemical reaction process that is being modelled from extrinsic noise, which comes from outside of the particular process [51]. In the model (3.1), extrinsic noise occurs when the parameters θ have different values for different cells in the population. In that case, the distribution of parameter values θ within the population can be modelled with a (constant) probability distribution. The characteristics of such a distribution are then independent variables of the model and can be subject to parameter estimation [52].

3.3. Models from population balance equations

3.3.1. Description of the model class

PBEs have been in use for several decades now as models for heterogeneous cell populations [13]. In these models, the state of the cell population is described by a time-dependent number density function $N(t, x)$, where t is the time and $x \in \mathbb{R}^n$ is the cell state as a vector of variables describing an individual cell, for example, cell size, cell age, protein expression levels, intracellular metabolite concentrations or other cellular variables. In models where spatial position plays a role, x may also contain cell position information. Integrating the number density function over a given region $\Omega \subset \mathbb{R}^n$ gives the number of cells for which the cell state x lies within Ω. The integral of the number density function over the complete state space gives the total number of cells in the population. The number density function is illustrated in figure 2. Note that the number density function is a continuous approximation that is appropriate for large cell populations.

A prototypical basic PBE is given by the partial integro-differential equation

$$\begin{aligned}\frac{\partial N}{\partial t}(t, x) + \mathrm{div}(f(x)N(t, x)) &= -d(x)N(t, x) - b(x)N(t, x)\\ &\quad + 2\int_{\Omega} b(\xi)\varphi(x, \xi)N(t, \xi)\,\mathrm{d}\xi.\end{aligned} \tag{3.5}$$

Here, the vector field $f(x) \in \mathbb{R}^n$ describes the temporal evolution of the cell state x, in the sense that dynamics of the cell state are modelled by the differential equation

$$\dot{x} = f(x). \tag{3.6}$$

Among the terms on the PBE's right-hand side, the scalar function $d(x)$ describes the rate of cell removal through cell death or an outflow from the system, and $b(x)$ describes the birth rate for new cells based on cell division. The term $-\,b(x)\,N(t, x)$ describes the removal of the mother cell upon cell division, while the integral term $\int_{\Omega} 2b(\xi)\varphi(x, \xi)N(t, \xi)\mathrm{d}\xi$ captures the appearance of the two daughter cells. The functions b and d typically depend on the cell state x. The scalar function $\varphi(x, \xi)$ is called the partition kernel and describes the probability that a daughter cell starts in cell state x after division, given that the mother cell had state ξ. Figure 3 illustrates how the cellular dynamics and cell division influence the dynamics of the number density function $N(t, x)$.

To solve the PBE (3.5), an initial condition and boundary conditions are needed. The initial condition $N_0(x) = N(0, x)$ is a density function over the cell state x. Which type of boundary conditions needs to be used depends on the considered domain Ω and the cellular dynamics $f(x)$. One common type is a no-flux condition, i.e. individual cells do not move their state across the boundary of Ω, which is represented by $\mathbf{n}(\bar{x}) \cdot f(\bar{x})N(t, \bar{x}) = 0$ for $\bar{x}$ on the corresponding boundary and $\mathbf{n}(\bar{x})$ the normal vector of the boundary at $\bar{x}$. Otherwise, boundary conditions will usually be derived from a cell 'influx' on one side of the domain, for example, when using the cell age as cellular variable [5], which starts at 0 for all newly born cells.

In some cases, only the change in the cell states x within the population is being modelled, while the cell birth and/or the cell death processes can be neglected in the model. This is usually appropriate for regulatory or signalling systems if they are studied on short time scales. In these models, the right-hand side of the PBE (3.5) will be set to zero, and the only relevant mechanism is the cell internal dynamics described by $f(x)$, which leads to a change in the number density function $N(t, x)$ through the divergence term $\mathrm{div}(f(x)N(t, x))$.

In this exposition, stochasticity in the cellular dynamics is not considered. Nevertheless, it is relatively straightforward to include that in PBE models, using for example a stochastic differential equation instead of the deterministic cellular dynamics (3.6). In the PBE (3.5), this would lead to the occurrence of a diffusion term representing the stochastic effects [53], which is equivalent to the diffusion term in the Fokker–Planck equation (3.4), if the chemical Langevin equation is used for the intracellular dynamics [49].

Regarding the numerical solution of the PBE, two main approaches can be distinguished. One approach uses a finite-dimensional approximation of the number density function $N(t, x)$ over the domain of the cell state x, either with a grid or as a series expansion. The grid-based approximation is then combined with finite difference [54] or finite-element [55] methods, resulting in a high-dimensional linear differential equation similar to the finite-state projection for CME models (3.2). More recent methods take special care to reproduce moments of the number density functions up to a certain order without approximation error [56]. The series expansion in turn is used with spectral methods for the discretization of the PBE [57,58]. Owing to the curse of dimensionality for grid-based methods, these are mostly useful for low-dimensional cell states. Special hierarchical simulation schemes may allow one to solve models of higher dimensions, where in a case study a six-dimensional cell state has been considered [59].

The other approach is based on solving the cell state dynamics (3.6) directly. For example, solutions based on the method of characteristics [60,61] fall into this category. Also stochastic, Monte Carlo-type methods [62,63] are based on this approach. Based on simulated samples of individual cells, density estimation techniques [64] are then used to reconstruct the density function $N(t, x)$. This approach can also be combined with the grid-based method in the so-called hybrid simulation methods [65,66].

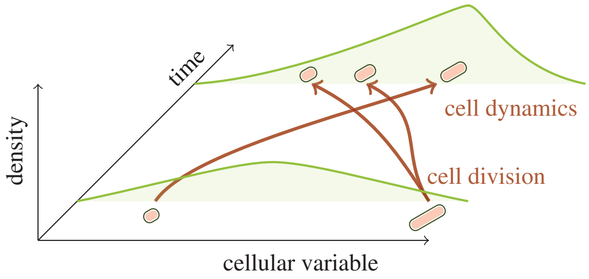

Figure 3. Illustration of the effect that cell division and cellular dynamics have on the number density function as represented in PBE models. Number density functions at different time points are represented by green curves. Pointed red curves illustrate the dynamics of individual cells in the state space.

When parameters are considered that remain constant over time for a single cell, but may be heterogeneous within the population, the cell state x needs to be defined as the combination of dynamic variables z and constant parameters p:

$$x = \begin{pmatrix} z \\ p \end{pmatrix}, \tag{3.7}$$

and the dynamics can be split into

$$\dot{z} = f_z(z, p) \quad \text{and} \quad \dot{p} = 0, \tag{3.8}$$

where f_z is a parameter-dependent vector field describing the change of the dynamic variables z [53]. The vector field f in the PBE (3.5) and (3.6) is then given by

$$f(x) = \begin{pmatrix} f_z(z, p) \\ 0 \end{pmatrix}. \tag{3.9}$$

In this case, the number density function $N(t, x)$ describes the joint density over the dynamic variables and the constant parameters.

The PBE can also include constant parameters where the same value applies to all cells in the population. Examples are kinetic rate constants based on substrate properties, sequence-based parameters for a clonal population, and population level or environmental parameters. Such parameters typically occur in the cellular dynamics $f(x)$, the death rate $d(x)$, the birth rate $b(x)$ and the partition kernel $\varphi(x, \xi)$. These are also called the *intrinsic physiological state* (IPS) functions.

In the context of population models such as the PBE (3.5), measurements to be used for estimation need to capture at least part of the cell population heterogeneity modelled with the number density function $N(t, x)$. A typical set-up is to measure a single or a few relevant variables out of the cell state x. The variables that can be measured from individual cells in this way include morphological variables such as the cell size and the abundances of any molecules for which a detectable marker is available, for example, using fluorescent staining. Measurements are then taken from many cells from the population, using, for example, high-throughput microscopic imaging or flow cytometry. Mathematically, we describe such a measurement for a single cell by an output equation

$$y = h(x), \tag{3.10}$$

where $y \in \mathbb{R}^m$ is the measured variable (or vector of measured variables), also called the output, and the function h describes how the measurement depends on the cell state x. Combining the measurements from a large number of cells at the same time point t results in an output number density function $N_{y(t)}(y)$ over the space of measured variables. The output number density function $N_{y(t)}$ is related to the state number density function N for the cell state through the integral relation

$$\int_{B_y} N_{y(t)}(y)\,\mathrm{d}y = \int_{h^{-1}(B_y)} N(t, x)\,\mathrm{d}x, \tag{3.11}$$

which holds for any subset B_y of the measurement space $\mathbb{R}^m$ and gives the number of cells for which the output y is in B_y, or equivalently for which the state x is in the pre-image $h^{-1}(B_y)$.

When a measurement noise model is used for the estimation, the resulting perturbation of the output density function $N_{y(t)}$ is obtained by a convolution with the noise probability distribution [7].

3.3.2. Related estimation problems

With the PBE models described in §3.3.1, the initial configuration at time 0 is given by the number density function $N_0(x)$. As discussed above, the variable x may contain parameters that are heterogeneous over the cell population, and such a heterogeneity would thus be a part of the initial configuration. Model parameters are either the IPS functions themselves or real-valued parameters that characterize these functions. According to the distinction between state and parameter estimation, estimation problems for PBE models will thus target the initial cell number density function $N_0(x)$ to be estimated [7,53,67,68], or the IPS functions such as division or death rates [69–71]. Both problems are in principle non-parametric estimation problems, since functions have to be estimated, but can be translated into real-valued estimation problems by a suitable parametrization of the target functions.

Estimation problems in PBE models are typically based on data in the form of population snapshots on the measured variable y according to the output equation (3.10), which is used to construct the output number density functions $N_{y(t)}$ (3.11) at different points in time. In the multi-dimensional measurement cases, the variables would ideally be measured simultaneously from the single cell, so that the full multi-dimensional output density function can be determined and correlations among measured variables will be known [72–74]. However, if the different variables are measured in separate experiments, only a one-dimensional output density for each variable, and not the full output density function may be available [75].

To summarize, a typical estimation problem for PBE models uses a time sequence of measured output density functions $N_{y(t)}$ to determine either the full-state density $N_0(x)$ (observation problem) or the IPS functions (parameter estimation problem) such that the model (3.5), parametrized according to the estimation result, best fits the available measurements.

Another common way to construct estimation problems in PBE models is by using the moments of the number density function as measurements. Moments are related to the number density function through the integral equation

$$M_\alpha(t) = \int_\Omega x^\alpha N(t, x)\,\mathrm{d}x, \tag{3.12}$$

where α is the order of the moment, potentially a multi-index for higher orders. While first-order moments, which are in principle average values, are available from simple bulk population measurements, higher-order moments are usually inferred from single-cell measurements according to (3.12). If not all state variables are measured, the moments for the output variables y would be obtained by integrating over the output density function $N_{y(t)}$ (3.11) instead. Moments are typically used for the estimation of IPS functions, though it would also be possible to estimate moments of the full-state density function from moments of the output.

3.4. Structured cell population models

3.4.1. Description of the model class

While PBE models such as discussed above consider the heterogeneity to be fully on a continuous scale, the so-called structured cell population models distinguish between different subgroups of cells by assigning individual cells to specific subpopulations. Such subpopulations may be defined by the cell type in a model for cell differentiation, the cell's generation as characterized by the number of divisions it has gone through in a growing population, or discrete variables such as copy number of a plasmid in the cell [76]. In systems biology research, structured cell population models have been used to model immune cell populations [77], stem cell proliferation [78] and differentiation [79,80], or ovarian follicle development [81].

Classically, structured cell population models use the association of each cell to one of multiple subpopulations as the only aspect of heterogeneity in the model. In that case, the full state of the population is described by the number of cells in each subpopulation. By a continuous approximation, one can describe this state with a real-numbered vector $N \in \mathbb{R}^k$, where the dimension k is equal to the number of subpopulations. The resulting dynamic model is then a differential equation of the form

$$\dot{N}(t) = AN(t) + B(t), \tag{3.13}$$

where A is a matrix that describes the transition rates from one subpopulation to another one, as well as population changes within one subpopulation through cell division and death or removal. B may describe a constant source of new cells entering the population; not through cell division, but from external sources. Such models are easily solved by standard differential equation solvers. In the model variant (3.13), the transition, birth and death rates going into the matrix A are constant. Then, one obtains a system of linear differential equations as cell population model, which can even be solved analytically. Alternatively, the transition rates might depend on the population sizes, and the model would use the term $A(N(t))$ instead, or they could be formulated as a time-varying expression $A(t)$. Sometimes a time delay is included with some of the state variables $N(t)$ in the right-hand side [81,82].

In some systems, the heterogeneity in the population is not only due to the existence of multiple subpopulations, where each cell can be associated with one subpopulation, but also includes continuously distributed variables, where a real-valued state variable (or vector) x is used to describe the properties of each individual cell. For these cases, combinations of the simple structured models as in (3.13) with the PBE models as described in section §3.3 have recently been proposed in a number of studies. These models basically consist of one PBE of the form (3.5) for each of the subpopulations [76,83,84]. As a simple example, consider the model structure

$$\frac{\partial N_i}{\partial t}(t, x) + \mathrm{div}(f_i(x)N_i(t, x)) = -\left(d_i(t) + \sum_{j=1}^{k} a_{j,i}(t)\right)N_i(t, x) + \sum_{j=1}^{k} a_{i,j}(t)N_j(t, x). \quad (3.14)$$

Note that the left-hand side is, as in the basic PBE (3.5), predominantly based on the dynamics f_i of the cell state variable x. These clearly may be different from one subpopulation to the other. Of course, cell division within each subpopulation can be considered as well, which would be done by adding the corresponding terms from the right-hand side of the basic PBE (3.5) to equation (3.14) for each subpopulation. In addition to the terms for cell division and removal, these models typically include transitions of cells from one subpopulation to another, modelling, for example, division events when subpopulations are structured according to generations, or cell differentiation events for a model including several cell types. In (3.14), these transitions are modelled by the terms with the transition rate $a_{i,j}(t)$, describing the transitions from subpopulation j to subpopulation i.

3.4.2. Related estimation problems

Estimation problems for structured cell population models need to use measurement of subpopulation sizes as data in the case of the simple model (3.13), or histogram data for a model with multiple PBEs as in (3.14). A frequent experimental restriction that further complicates the estimation is that subpopulations cannot be separated easily before performing a measurement. Thus, if, for example, the single-cell expression levels of specific markers are measured, a combined distribution from all subpopulations will be obtained as measurement data. It may be straightforward to distinguish subpopulations afterwards, if, for example, a marker is present in one subpopulation but not the other, but sometimes it is more involved and constitutes a relevant estimation problem to quantify the subpopulations.

Relevant properties to be estimated in structured cell population models are the initial configuration or model paramaters. The former is determined by the initial states of the subpopulations, either just their size $N_i(0)$ in the basic structured models (3.13), or the number density function $N_i(0, x)$ for each subpopulation in the hybrid models (3.14). Model parameters will be the transition rates and source terms written in the matrix A and vector B in the basic model (3.13), and for the hybrid model (3.14) accordingly the transition and death rates a_i and d_i, and in addition the cell state dynamics f_i.

3.5. Individual-based cell population models

3.5.1. Description of the model class

Individual-based models (or agent-based models) are used in a wide range of scientific disciplines for the simulation of large systems that are composed of moderately to highly complex individual entities, named the agents. In the context of cell populations, it means that a sufficiently large number of cells is being modelled individually. The behaviour of each cell is described by a dynamic model, and additional model elements capture the interactions among cells and between cells and the environment. Situations where individual-based models for a cell population are frequently used in systems biology include spatial organization and dynamics in the extracellular environment, for example, in tumour development [85,86], or interactions among cells through diffusive exchange with the extracellular environment, for example, in the synchronization of oscillations [87–89].

The class of cell ensemble models, reviewed in [15], is a particular instance of individual-based models with reduced complexity. Thereby, the cellular models only describe intracellular dynamics arising, for example, from metabolism or biochemical signal transduction and exchange reactions with the environment, and are typically formulated as differential equation models. Moreover, there are no direct interactions among individual cells, but only interactions via a joint extracellular medium, usually assumed to be well mixed. This is often an appropriate approximation in, for example, bioreactors or in the blood stream, cell colonies where the nutrient availability is not too much affected by the spatial organization of cells, or the analysis of signalling systems which are not affected by the population's spatial organization. The simple structure of these models allows for a very efficient computational simulation: a recent software implementation described in [90] is able to simulate a population model of 10 000 cells coupled to an extracellular environment, modelled by 22 differential equations each, over a simulated time of 30 min, in the quite reasonable computation time of about 5 h.

Instead of using differential equations, individual cells may also be modelled stochastically. In that case, sample trajectories for individual cells are generated through a stochastic simulation. An example for this is the variable number Monte Carlo simulation proposed by Mantzaris [63].

Cell division or death have been considered by copying or removing individual cells [5], although these population dynamics may be neglected if the model considers only a short time range compared with the cell cycle duration, or if the model is not meant to describe the population size accurately [87,91].

Individual-based models are also frequently used to generate full cell genealogies, which are tree structures that contain information about relevant cellular variables over the individual cell's lifespan as well as mother–daughter relations [92]. Such genealogies, for example, allow one to determine the timing from intracellular differentiation decisions to an observable outcome [42]. Looking at the trajectories of individual cells in the genealogy also permits one to clarify the correlations in cellular variables such as protein amounts through cellular ancestry relations [5].

3.5.2. Related estimation problems

Individual-based models are most useful when single-cell trajectory data are available. While estimating the individual states from such trajectories is in principle possible, it is often not the prime interest, because in a large population an individual cell does not matter that much. Instead of estimating state or parameter values for individual cells, one is more often interested in population-wide parameters that are involved in the cell dynamics or interactions. For example, cell division may be modelled as a stochastic process for the individual cell, where the probability of division depends on the cell state in combination with a population-level parameter such as a division rate constant, and the estimation would target this rate constant. For variables specific to a single cell, instead of estimating them individually, it is often more useful to estimate probability distributions that describe the distribution of such parameters within the population.

Cell ensemble models can also be formulated as mixed-effects models [93], for which efficient estimation methods can be used when trajectory data for individuals are available.

3.6. Discussion

There is quite a range of model classes available to describe and simulate heterogeneous cell populations. In principle, the choice of which class to use should be based on the objectives pursued with the modelling, and the experimental methods should be chosen such that they provide appropriate data to parametrize the model. In practice, one however often has the situation that the experimental methods and available measurements are previously determined, and one has to select a modelling approach that works with the resulting data. Fortunately, the model frameworks presented here are closely linked to each other, so that in most cases it is easy to transition from one framework to the other. For example, if the intracellular dynamics are represented by a chemical reaction network and no cell death or division is accounted for, the Fokker–Planck equation (3.4) will be the same as the PBE (3.5) with the stochastic diffusion term added. Similarly, an individual-based population model can be interpreted as a sampling-based approximation of the PBE, and if the underlying cellular dynamics are formulated as a stochastic chemical reaction network, it is basically the same as simulating a chemical reaction network with the Gillespie algorithm [94].

Some specific limitations and advantages of the available model classes have to be taken into account when choosing which one to work with. Points that need to be clarified are, for example, how relevant intrinsic noise is to the process under study, or whether the overall timescale that is to be modelled warrants the inclusion of population dynamics in the model, i.e. cell division and death. Intrinsic noise will have a larger effect if low numbers of molecules are involved, for example, due to on/off regulation of a gene. In that case, the model should include single-cell stochastics, either based on the CME, by including a stochastic term in the PBE models, or by a stochastic simulation in the individual-based models. With CME-based models, including population dynamics is difficult, since changes in the overall cell number are not directly captured by the probability density function that is used there. Instead, it is then more appropriate to use one of the other model classes, i.e. either the PBE models with a number density function, or the individual-based models where changes in cell numbers can be tracked directly.

The different model classes also have different limitations regarding the dimension up to which they can be simulated efficiently. The dimension is limited most severely when a full probability distribution (for CME models) or density function (for PBE models) is to be computed. Otherwise, using an appropriate approximation or sampling technique, the computational effort can be similar to that of solving an ODE.

4. Estimation methods

4.1. Identifiability and observability analysis

A key question in any estimation task is whether the given measurements are informative enough to infer values of the underlying states or parameters. For parameter estimation, this is called identifiability analysis, while for state estimation it is called observability analysis, even though the concept and methodology overlap to some extent.

With classical ODE models, the questions of identifiability and observability are well characterized in systems biology [95]. Observability analysis and some methods of identifiability analysis are commonly performed on the model alone [22,96,97]. With likelihood-based identification methods as introduced in §4.3, it is more common to just evaluate the uncertainty after the estimation has been performed, for example, in the form of confidence intervals, and conclude about identifiability problems from that. A systematic approach to do that is for example with profile likelihoods [98], which also permit one to distinguish between structural and practical non-identifiability. While analytical identifiability analysis methods give more insight into potential reasons for non-identifiability, they are typically only feasible for low-dimensional problems, whereas likelihood-based methods also work well in higher dimensions.

In the case of population balance models, the term ‘identifiability’ has already been used in [60], but referred very specifically to a direct computational reconstruction of IPS functions by size measurements from the population in the case of balanced growth. More generic methods for identifiability and observability analysis are just emerging. As with classical models, an uncertainty analysis with a Bayesian approach yields confidence regions [7], which in principle could be used to detect potential identifiability problems [74]. A more formal identifiability and observability analysis for population balance models, though considering only intracellular dynamics, not cell division and death, has been developed in a recent series of papers [67,99,100]. In these papers, the problem of estimating heterogeneous parameters in the form of a density function or the initial state density function $N_0(x)$ is considered. It was proven mathematically that a necessary condition for the population model to be observable or identifiable is that the underlying single-cell dynamics (3.6) with the corresponding single-cell output (3.10) already has that property. Based on an approach borrowed from computed tomography [100], a sufficient condition has been determined for the specific case where the single-cell dynamics $f(x)$ are linear. In that case, the population model is observable, if the single-cell model is

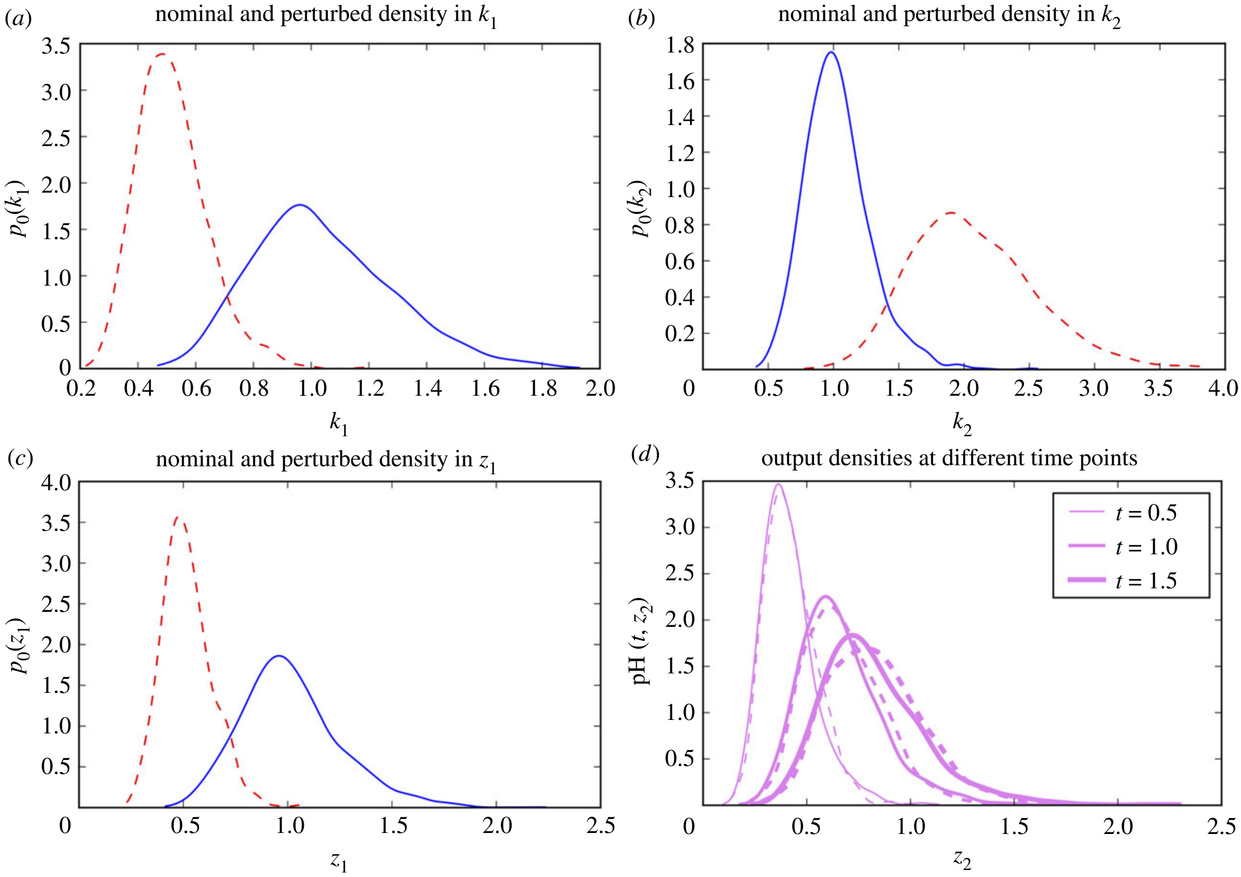

Figure 4. Density functions in a non-identifiable example system. (*a–c*) Marginal initial-time density functions of different cellular variables with the full line the nominal case and the dashed line the perturbed case. (*d*) Density functions for the measured variable are indistinguishable in the nominal (solid line) and perturbed (dashed line) case, even when taking multiple time points into account (thickness of the lines). Adapted from [99]. (Online version in colour.)

observable and the output vector y has dimension $n-1$, where n is the dimension of x [67]. This means that all but one variables need to be measured in order to reconstruct the full underlying density function, which appears to be a severe restriction for practical estimation applications, where one could have a single-cell model with dozens of state variables, but only a few can be measured simultaneously from single cells. Unfortunately, even though the observability condition is only sufficient, the limitation is real in actual models: an example with a three-dimensional model is constructed where only a single variable is measurable, and while the single-cell model itself is observable through that measurement, the corresponding population model is not [67]. In fact, some of the second-order moments in this model cannot be determined uniquely from the measured output distributions. While models for intracellular dynamics in systems biology are mostly nonlinear, the condition for linear models thus already shows the limitations inherent to the observability problem for the population. Even though the problem for a general nonlinear system is still unsolved, one can unfortunately not expect that the conditions for the general case will be less restrictive than for a linear system.

The issue of non-identifiability is illustrated in figure 4 with results from an example discussed in [99]. This is an example where even the single-cell dynamics are not identifiable. In that case, if the density functions of the cellular variables are changed according to the non-identifiability, the density functions of the measured variables are not affected at any point in time.

4.2. Convex optimization for density functions

In PBE models as described in §3.3, the problem of estimating a density function for the single-cell states or extended states defined in (3.7) can be formulated as a convex optimization problem. The initial density function $N_0(x)$ to be estimated is written as a linear combination of ansatz densities

$$N_0(x) = \sum_{i=1}^{k} c_i \varphi^{(i)}(x), \tag{4.1}$$

with coefficients c_i to be determined through the estimation and predefined ansatz functions $\varphi^{(i)}(x)$. In past studies, the so-called hat functions [53] and Gaussian kernels [7,68] have been used for the ansatz functions. For each of the ansatz functions, the PBE is solved with the ansatz function as the initial density, and the corresponding output densities $\varphi^{(i)}_{y(t)}(y)$ are obtained according to the output equation (3.11). Because the PBE is a linear equation, for a linear combination of ansatz functions as in (4.1), the output is the same linear combination of the individual output densities [53]. Thus, in order to reconstruct the initial density N_0, one can construct an optimization problem to bring the model's output density as close as possible to a measured output density $N_{y(t)}$:

$$\min_{c_i} \left\| N_{y(t)} - \sum_{i=1}^{k} c_i \varphi^{(i)}_{y(t)} \right\|^2. \tag{4.2}$$

In past studies, the L^2 norm has been used for this optimization [53,68], though using other norms is also possible.

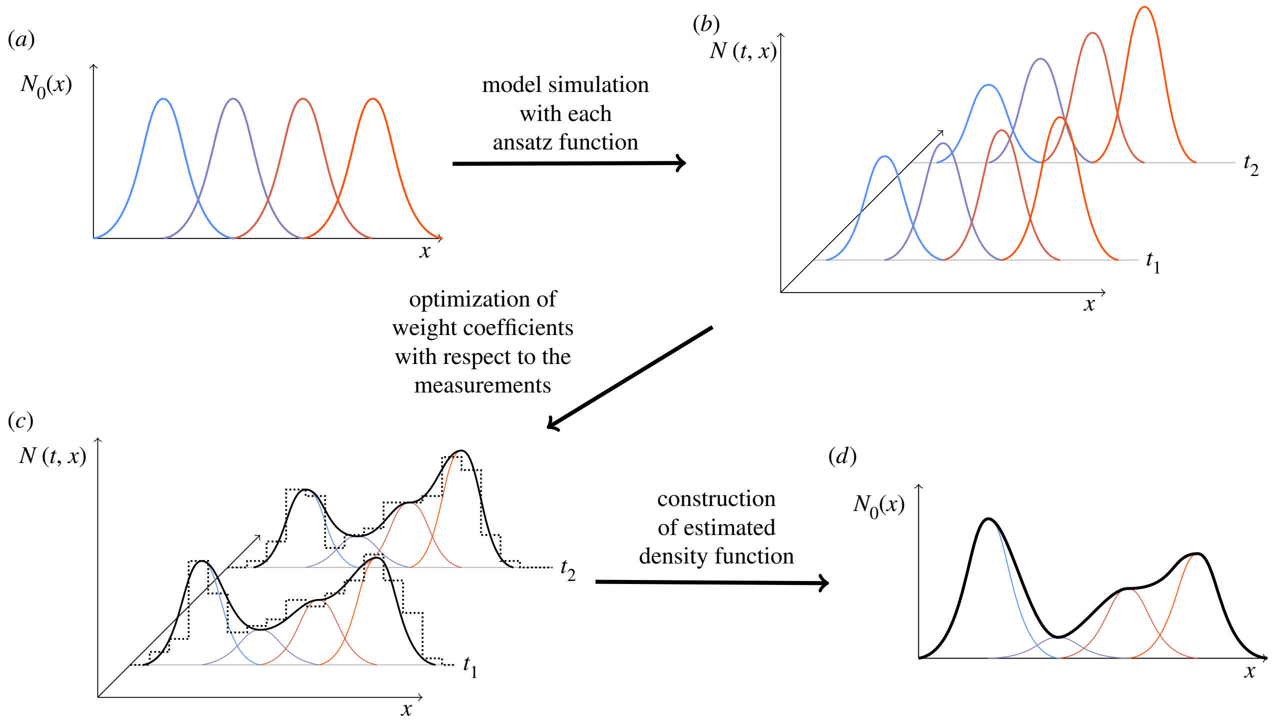
(a)
$N_0(x)$
x
model simulation with each ansatz function
(b)
$N(t, x)$
t_2
t_1
x
optimization of weight coefficients with respect to the measurements
(c)
$N(t, x)$
t_2
t_1
x
construction of estimated density function
(d)
$N_0(x)$
x

Figure 5. Estimation procedure using ansatz functions and a convex optimization approach. (*a*) Ansatz functions φ_i. (*b*) The population model is simulated for each ansatz function and the corresponding output density at each measurement time point is obtained. (*c*) The coefficients c_i are optimized to minimize the overall deviation of the simulated output density functions to the measured output density functions (dotted line). (*d*) The estimation result is constructed by combining the ansatz functions with the optimal weight coefficients c_i.

Another measure to compare density functions from the model and the data is the Kullback–Leibler divergence, which has been used in a recently proposed particle filter for the estimation of cell populations [101].

The PBE in (4.2) needs to be solved k times before running the optimization, but not within the optimization itself. After a suitable discretization of the involved density functions, this is a simple quadratic optimization problem in the coefficients c_i that can be solved efficiently to a global optimum. An overview of this estimation procedure is shown in figure 5. In order to avoid overfitting, an alternative approach that has been proposed is to put an upper constraint on the model's deviation from the experimental data, while maximizing the entropy of the estimated density functions [102]. A maximization of the density function's entropy based on histogram data has also been considered in [103]. The maximization of entropy is motivated by the goal to keep the estimated density function as broad as possible while maintaining the model error beneath an upper limit.

As discovered by Zeng *et al.* [100] through the observability analysis, the estimation of density functions in heterogeneous cell population models is essentially a mathematical tomography problem. Therefore, reconstruction methods that have been developed in the context of computed tomography can directly be transferred to the estimation of density functions in cell populations. A classical reconstruction method in tomography is the filtered back projection (FBP), which is based on an analytical inversion of the Radon transformation. The Radon transformation is the classical description in tomography of how the density function for the states is mapped to the density function for the outputs, according to the output equation (3.11). However, the FBP is quite sensitive to the output equation not covering all possible projection directions, which typically occurs in population models because the system dynamics do not give a sufficiently large 'rotation' of the density function as would be required for the classical tomographic reconstruction. In that case, the reconstruction from the FBP suffers from streak artefacts [68]. A more reliable reconstruction method from tomography is the algebraic reconstruction technique, which is based on a grid discretization of the state space and the approximation of the output equation by a finite-dimensional linear mapping. The reconstruction then amounts to finding a minimum-error norm inversion of this mapping, commonly via an iterative method such as the Landweber algorithm [104]. The result is a gridwise reconstruction of the initial state density in the model corresponding to the measured output densities [68].

4.3. Likelihood-based estimation methods

Model elements on the level of the complete population, such as birth or death rates in the PBE (3.5) or the transition rates in structured population models, are often parametrized by real-valued parameters. A reliable approach to estimate such parameters from measured data are likelihood-based estimation methods using Bayes' theorem. Thereby, one searches for a parameter value θ^* that, based on observations $\mathcal{Y}$, maximizes the posterior probability distribution

$$p(\theta|\mathcal{Y}) = \frac{p(\mathcal{Y}|\theta)\, p(\theta)}{p(\mathcal{Y})}. \tag{4.3}$$

Here, $p(\mathcal{Y} \mid \theta)$ is the conditional probability (or likelihood) of obtaining the measurement $\mathcal{Y}$ given θ as parameter values, $p(\theta)$ is the prior distribution collecting *a priori* knowledge about parameter values, and $p(\mathcal{Y})$ is to be used as a normalization constant. Estimation methods typically rely on either an optimization with the posterior distribution $p(\theta \mid \mathcal{Y})$ as objective function, or the direct determination of this distribution as a function of θ.

Since local optima occur frequently in the posterior distribution, multi-start local methods [105] or global stochastic/hybrid methods [106,107] are commonly used for the optimization. An alternative are probabilistic methods such as Markov chain Monte Carlo [43,52] or sequential Monte Carlo/particle filtering [42]. These methods generate a collection of particles representing different estimates together with their posterior distributions, and from this collection, an overall estimate can be generated by either considering the most probable particle (maximum likelihood), or by computing the expected parameter values as an average of all particles. Commonly used methods for the direct determination of the likelihood include Markov chain Monte Carlo sampling or profile likelihoods [98].

A range of studies has been using the likelihood-based approach for the estimation of population-level parameters such as transition rates between different subpopulations or statistical parameters (such as mean, variance) describing the distribution of heterogeneous parameters within the population, using CME-based models [52] as described in §3.2 or individual-based models [42,43] as described in §3.5. In these cases, the likelihood is obtained as the product of the individual measurements' likelihoods over both observation time points and sampled cells. These methods are also able to include single-cell time courses and lineage data in the form of mother–daughter relationships between cells. Thereby, an iterative calculation is applied to compute the likelihood of an observed lineage tree from parameters on the population and single-cell level with the help of an individual-based model [43]. Owing to the computational complexity of the likelihood calculation for population models, current methods are typically only feasible for less than about 20 parameters in total. That works well for simple models of single-cell dynamics, but more complex models appear to be out of scope currently, unless most parameters are assumed to be known *a priori*. Recent case studies considered the estimation of four to seven parameters [42,43] with lineage tree data. With snapshot data, the likelihood can be computed more efficiently, allowing for more parameters to be estimated—for example, [52] considered a model with 15 parameters that were estimated from snapshot data using moment dynamics.

The likelihood-based approach has also been used with population models for the estimation of a density function for parameters that are heterogeneous within the cell population [7,108]. In this case, the parameter θ is actually a density function, and the likelihood $p(\mathcal{Y}\,|\,\theta)$ can be evaluated by multiplication and integration over snapshot data from individual cells as

$$p(\mathcal{Y}\,|\,\theta) \sim \prod_{y^{(i)}\in\mathcal{Y}} \int_{\mathbb{R}} p(y^{(i)}\,|\,\eta)\theta(\eta)\,d\eta, \tag{4.4}$$

where $y^{(i)}$ is the snapshot measurement from cell i, and η the heterogeneous parameter. This integral is efficiently evaluated by a pre-computation with appropriate basis functions and Monte Carlo integration. In [7], a one-dimensional and a two-dimensional density function for a heterogeneous cellular parameter were approximated with 15 and 100 basis functions, respectively. Using Markov chain Monte Carlo approximation, the posterior probability distributions for the corresponding coefficients were then determined and evaluated with respect to the maximum-likelihood estimate and confidence intervals.

4.4. Estimation of model functions and intrinsic physiological state functions

For PBE models, the estimation of IPS functions from data has been an active research area since these models have been established. Even before the formal development of PBEs, early studies considered the estimation of single-cell growth rates, cell division rates and the daughter size distribution for populations of exponentially growing bacteria [109]. With the establishment of PBEs as a formal description of population dynamics, this estimation problem could also be treated more formally as an inverse problem [14, ch. 6]. Thereby, it is typically assumed that the full-state density function $N(t,x)$ is available as measurement at multiple points in time. While this assumption is realistic if the single-cell state x is only the cell size, as in classical PBE models for cell populations, the approach will need to be refined for more general cases. In relation to the PBE model (3.5), the IPS functions typically considered are the cell division rate $b(x)$, the partition kernel $\varphi(x, \xi)$ and the single-cell dynamics $f(x)$ (which is the single-cell growth rate when x is the cell size). At this point, it is however not clear when it is possible to reliably estimate all IPS functions from measurements of only the density function $N(t, x)$.

Estimation methods developed in applications to particle systems have considered the reconstruction of only a subset of IPS functions from measurements of the density function. One example is the estimation of the breakage (or cell division) rate from the number density function, which has been solved as an inverse problem by introducing suitable basis functions, leading to a least-squares problem with a Tikhonov regularization [110]. Thereby, particle growth dynamics, corresponding to the intracellular dynamics $f(x)$ in the cell population case, have been neglected. In another approach, Mahoney *et al.* [111] have considered the estimation of the growth rate $f(x)$ from measurements of the particle size number density function over time as an inverse problem. The particle breakage rate and the partition kernel are neglected or need to be known in this approach. By contrast, the inverse problem studied in [112] is about the estimation of the cell division rate, while the single-cell dynamics are assumed to be known and the division is considered to be symmetric.

Past approaches specific to cell populations have based the inverse problem not only on the overall density function $N(t, x)$, with x the cell size, but also on the size distributions of the dividing cells and the new-born cells. In particular, Collins & Richmond [109] derived a closed-form expression for the single-cell dynamics $f(x)$ as a single-cell growth rate, and for the cell division rate $b(x)$. In later studies, a practically feasible approach was developed to experimentally determine the required three density functions $N(t,x)$ together with the size distributions of dividing and new-born cells, based on a sophisticated image analysis methodology [69]. The key limitation here is that the image analysis is restricted to rod-shaped cells, so the approach is only applicable to appropriately shaped species. Using a non-parametric estimation method, the thus constructed density functions were used to estimate the single-cell dynamics $f(x)$ in combination with the cell division rate $b(x)$ and the partition kernel $\varphi(x,\ \xi)$ [70]. Besides the closed-form expressions as in [109] for the single-cell growth and division rates, a non-parametric estimation with basis functions formulated as regularized least-squares problem has been used for the partition kernel.

Recently, the estimation of IPS functions using only dynamic measurements of the number density function $N(t,\ x)$ was studied in [71]. Moment equations were derived that relate the moment dynamics to the number density function in combination with two of the IPS functions, the cell division rate $b(x)$ and the single-cell growth rate $f(x)$. From measurements of the number density function, the corresponding moment dynamics can be inferred and used to construct an inverse problem for these IPS functions in the form of a standard least-squares problem. The partition kernel $\varphi(x,\ \xi)$ is not estimated, but does not need to be known either for the estimation of the other functions. However, the problem turned out to be quite ill-conditioned, and strong regularizing constraints need to be made in order to get useful estimates.

4.5. Estimation of population dynamics with persistent cell labelling

Persistent cell labelling with CFSE or BrdU as discussed in §2, together with the structured cell population models described in §3.4, permits the estimation of relevant population parameters such as cell division and death rates.

Early studies with CFSE as a label used relatively simple structured cell population models of the type (3.13) [113,114]. Later it was realized that these models allow only a weak fit to data compared to more complex models [115].

One model extension that was subsequently developed is the label-structured population model, which is a PBE similar to (3.5), where the intracellular variable x represents the intensity of the CFSE label in the cell [116]. This model was then used to estimate a cell birth rate $b(x)$ as a function of the label intensity as well as some scalar model parameters including a constant death rate d from CFSE labelling data of a T lymphocyte population [117]. Thereby, the birth rate was approximated by splines, and the estimation was performed by least-squares optimization coming from a maximum-likelihood approach to minimize the model-data deviation for the number density function $N(t,\ x)$. A detailed discussion of the relevant estimation techniques for this model type has also been published in [118].

In the label-structured population models, the dependency of the birth rate on the label intensity is more a phenomenological representation of an observed correlation than a realistic representation of the underlying biology [116]. Further model extensions thus generalized this to division and label-structured population (DLSP) models, where multiple discrete subpopulations according to the division number are combined with a heterogeneity in the label intensity x, similar to the structured population model (3.14) [84]. Based on this model type, an estimation of generation- and time-dependent birth and death rates as well as parameters related to the labelling such as label decay, the initial label distribution and autofluorescence was performed [119]. Using a Bayesian approach, an efficient computation of the likelihood function was proposed by approximating the number density function as a sum of lognormal densities, and the posterior distribution $p(\theta|\mathcal{Y})$ was evaluated with Markov chain Monte Carlo sampling to determine the maximum likelihood estimate and associated uncertainty. In a case study with the same data on T lymphocyte population dynamics as used in previous studies [116], it was shown that this model and estimation method yields a good fit with low uncertainty in the estimates (at least where it matters for the model's predictions).

In a recent extension to the DLSP model [120], the addition of the cell age into the cellular variables x was proposed. Thereby, each of the number density functions $N_i(t, x)$ in (3.14) becomes a time-varying density over a two-dimensional property space for x. With that model, different hypotheses about the birth rates $b_i(x)$ could be formulated, such as a dependency on division number, cell age, or both. For a model of T lymphocyte proliferation, 15 model alternatives were formulated where the functional parameters are discretized to yield 12–29 real-valued parameters per model [120]. For each of these models, the authors performed a maximum-likelihood estimation with stochastic and multi-start local optimization methods, where they observed better convergence property of the multi-start local methods compared to the stochastic methods. On the biological side, a key finding was that making birth rates dependent on cell age significantly improved the model's fit to data compared to constant or division number dependent rates.

Note that with the DLSP and subsequent models not the full system state can be measured: since only the label intensity is measurable for single cells, only an output number density function $N_{y(t)}$ (3.11) with a marginalization over division numbers and cell ages is available as data.

4.6. Discussion

Clearly, a considerable range of methods is already available to solve estimation problems for heterogeneous cell populations. However, the choice of method is to a large extent dictated by the available experimental data, the model one works with, and the properties of the model that are unknown and should be estimated. Nevertheless, an important aspect to consider is the computational cost incurred with different estimation methods. Generally, this has two aspects: the computational cost of solving the model for specific parameter values and initial conditions (the 'forward problem'), and the cost of the optimization itself, i.e. how often the model needs to be solved until a reliable estimate is obtained.

One property of likelihood-based methods with sampling is that each additional data point adds to the computation cost. Thus, these methods face more challenges with, for example, high numbers of measured cells and individual-based models. In that case, it may be more efficient to use models and methods that work with density functions or histograms.

Convex optimization-based methods have so far been applicable to rather low-dimensional systems, but are very efficient for PBE models with population snapshot data and histograms. However, so far these methods do not yet take the full population dynamics including cell division into account, having focused on the heterogeneity in the cellular dynamics. To estimate the cell division dynamics, more specific methods as reviewed in §4.4 and 4.5 need to be used.

5. Conclusion and outlook

Heterogeneity in cell populations has found increasing interest over the past decade, and the corresponding computational models contribute significantly to understanding these systems. Connecting experiments and models critically relies on state and parameter estimation methods, which need to be adapted to the types of experimental data and models that are available for cell populations.

For a successful state and parameter estimation, three components need to fit together: the experimental data, the model class used to describe the population, and the estimation algorithm that is used to carry out the estimation. Typical data that are currently being used for model-based estimation are population snapshots or cell trajectory data, both with a few measured variables per cell. That data type fits well to model classes such as population balance models or structured population models, which describe heterogeneity in a few to a dozen variables (state variables and parameters combined) per cell. Also computational methods as well as current computers are efficient enough to deal with this problem complexity.

An example where these components do not yet fit well together is single-cell 'omics' data, with hundreds to thousands of measured variables per cell and potential heterogeneity in these variables. Models of heterogeneous cell populations with that number of variables per cell have not yet been formulated. But even if such models were available, one can doubt that data from a few hundred cells, as is the current state of the art for many 'omics' type single-cell measurements, would be sufficient to quantitatively estimate heterogeneous properties in a reliable manner. Moreover, the complexity of such an estimation would likely also go beyond the current computational capabilities. Yet, with improvements in both the cell number in these experiments and computational efficiency of modelling and estimation methods for very high-dimensional problems, it will hopefully become possible to use these data in future model-based estimation studies.

Data accessibility. This article has no additional data.

Competing interests. I declare I have no competing interests.

Funding. I received no funding for this study.

Acknowledgements. I thank Armin Küper for helpful comments on the manuscript.

References

1. Altschuler SJ, Wu LF. 2010 Cellular heterogeneity: do differences make a difference? *Cell* **141**, 559 – 563. (doi:10.1016/j.cell.2010.04.033)
2. Fraser D, Kaern M. 2009 A chance at survival: gene expression noise and phenotypic diversification strategies. *Molec. Microbiol.* **71**, 1333 – 1340. (doi:10.1111/j.1365-2958.2009.06605.x)
3. Delvigne F, Zune Q, Lara AR, Al-Soud W, Sørensen SJ. 2014 Metabolic variability in bioprocessing: implications of microbial phenotypic heterogeneity. *Trends Biotechnol.* **32**, 608 – 616. (doi:10.1016/j.tibtech.2014.10.002)
4. Binder D, Drepper T, Jaeger K-E, Delvigne F, Wiechert W, Kohlheyer D, Grünberger A. 2017 Homogenizing bacterial cell factories: analysis and engineering of phenotypic heterogeneity. *Metab. Eng.* **42**, 145 – 156. (doi:10.1016/j.ymben.2017.06.009)
5. Imig D, Pollak N, Strecker T, Scheurich P, Allgöwer F, Waldherr S. 2015 An individual-based simulation framework for dynamic, heterogeneous cell populations during extrinsic stimulations. *J. Coupled Syst. Multiscale Dyn.* **3**, 122 – 134. (doi:10.1166/jcsmd.2015.1070)
6. Fritzsch FSO, Dusny C, Frick O, Schmid A. 2012 Single-cell analysis in biotechnology, systems biology, and biocatalysis. *Annu. Rev. Chem. Biomol. Eng.* **3**, 129 – 155. (doi:10.1146/annurev-chembioeng-062011-081056)
7. Hasenauer J, Waldherr S, Doszczak M, Radde N, Scheurich P, Allgöwer F. 2011 Identification of models of heterogeneous cell populations from population snapshot data. *BMC Bioinf.* **12**, 125. (doi:10.1186/1471-2105-12-125)
8. Hilsenbeck O *et al.* 2016 Software tools for single-cell tracking and quantification of cellular and molecular properties. *Nat. Biotechnol.* **34**, 703 – 706. (doi:10.1038/nbt.3626)
9. Kobel SA, Burri O, Griffa A, Girotra M, Seitz A, Lutolf MP. 2012 Automated analysis of single stem cells in microfluidic traps. *Lab Chip* **12**, 2843 – 2849. (doi:10.1039/C2LC40317J)
10. Van Kampen NG. 1981 *Stochastic processes in chemistry and physics*. Amsterdam, The Netherlands: North Holland.
11. Gillespie DT. 1992 A rigorous derivation of the chemical master equation. *Physica A* **188**, 404 – 425. (doi:10.1016/0378-4371(92)90283-V)
12. Wilkinson DJ. 2009 Stochastic modelling for quantitative description of heterogeneous biological systems. *Nat. Rev. Genet.* **10**, 122 – 133. (doi:10.1038/nrg2509)
13. Fredrickson AG, Ramkrishna D, Tsuchiya HM. 1967 Statistics and dynamics of procaryotic cell populations. *Math. Biosci.* **1**, 327 – 374. doi:10.1016/0025-5564(67)90008-9)
14. Ramkrishna D. 2000 *Population balances. Theory and applications to particulate systems in engineering*. San Diego, CA: Academic Press.
15. Henson MA. 2003 Dynamic modeling of microbial cell populations. *Curr. Opin. Biotechnol.* **14**, 460 – 467. (doi:10.1016/S0958-1669(03)00104-6)
16. Ramkrishna D, Singh MR. 2014 Population balance modeling: current status and future prospects. *Annu. Rev. Chem. Biomol. Eng.* **5**, 123 – 146. (doi:10.1146/annurev-chembioeng-060713-040241)
17. Domach MM, Shuler ML. 1984 A finite representation model for an asynchronous culture of *E. coli*. *Biotechnol. Bioeng.* **26**, 877 – 884. (doi:10.1002/bit.260260810)
18. Ataai MM, Shuler ML. 1985 Simulation of CFSTR through development of a mathematical model for anaerobic growth of *Escherichia coli* cell population. *Biotechnol. Bioeng.* **27**, 1051 – 1055. (doi:10.1002/bit.260270717)
19. Dochain D. 2003 State and parameter estimation in chemical and biochemical processes: a tutorial. *J. Proc. Contr.* **13**, 801 – 818. (doi:10.1016/S0959-1524(03)00026-X)
20. Lillacci G, Khammash M. 2010 Parameter estimation and model selection in computational biology. *PLoS Comput. Biol.* **6**, e1000696. (doi:10.1371/journal.pcbi.1000696)
21. Berthoumieux S, Brilli M, de Jong H, Kahn D, Cinquemani E. 2011 Identification of metabolic network models from incomplete high-throughput datasets. *Bioinformatics* **27**, i186 – i195. (doi:10.1093/bioinformatics/btr225)
22. Farina M, Findeisen R, Bullinger E, Bittanti S, Allgöwer F, Wellstead P. 2006 Results towards identifiability properties of biochemical reaction networks. In *Proc. 45th IEEE Conf. on Decision and Control, San Diego, CA, USA, 13 – 15 December 2006*, pp. 2104 – 2109. (doi:10.1109/CDC.2006.376925)
23. Byrne H, Drasdo D. 2008 Individual-based and continuum models of growing cell populations: a comparison. *J. Math. Biol.* **58**, 657 – 687. (doi:10.1007/s00285-008-0212-0)
24. Mandy FF, Bergeron M, Minkus T. 1995 Principles of flow cytometry. *Transfusion Sci.* **16**, 303 – 314. (doi:10.1016/0955-3886(95)90002-0)
25. George TC *et al.* 2006 Quantitative measurement of nuclear translocation events using similarity analysis of multispectral cellular images obtained in flow. *J. Immunol. Methods* **311**, 117 – 129. (doi:10.1016/j.jim.2006.01.018)
26. Frei AP, Bava F-A, Zunder ER, Hsieh EWY, Chen S-Y, Nolan GP, Federico Gherardini P. 2016 Highly

multiplexed simultaneous detection of RNAs and proteins in single cells. *Nat. Methods* **13**, 269–275. (doi:10.1038/nmeth.3742)

27. Albayrak C, Jordi CA, Zechner C, Lin J, Bichsel CA, Khammash M, Tay S. 2016 Digital quantification of proteins and mRNA in single mammalian cells. *Molec. Cell* **61**, 914–924. (doi:10.1016/j.molcel.2016.02.030)
28. Bonhoeffer S, Mohri H, Ho D, Perelson AS. 2000 Quantification of cell turnover kinetics using 5-bromo-2′-deoxyuridine. *J. Immunol.* **164**, 5049–5054. (doi:10.4049/jimmunol.164.10.5049)
29. Wang D, Bodovitz S. 2010 Single cell analysis: the new frontier in 'omics'. *Trends Biotechnol.* **28**, 281–290. (doi:10.1016/j.tibtech.2010.03.002)
30. Lin J-R, Fallahi-Sichani M, Chen J-Y, Sorger PK. 2016 Cyclic immunofluorescence (CycIF), a highly multiplexed method for single-cell imaging. *Curr. Protoc. Chem. Biol.* **8**, 251–264. (doi:10.1002/cpch.14)
31. Bodenmiller B *et al.* 2012 Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators. *Nat. Biotechnol.* **30**, 857–866. (doi:10.1038/nbt.2317)
32. Weaver WM, Tseng P, Kunze A, Masaeli M, Chung AJ, Dudani JS, Kittur H, Kulkarni RP, Di Carlo D. 2014 Advances in high-throughput single-cell microtechnologies. *Curr. Opin. Biotechnol.* **25**, 114–123. (doi:10.1016/j.copbio.2013.09.005)
33. Hashimshony T, Wagner F, Sher N, Yanai I. 2012 CEL-seq: single-cell RNA-seq by multiplexed linear amplification. *CellReports* **2**, 666–673. (doi:10.1016/j.celrep.2012.08.003)
34. Streets AM *et al.* 2014 Microfluidic single-cell whole-transcriptome sequencing. *Proc. Natl Acad. Sci. USA* **111**, 7048–7053. (doi:10.1073/pnas.1402030111)
35. van den Brink SC, Sage F, Vértesy Á, Spanjaard B, Peterson-Maduro J, Baron CS, Robin C, van Oudenaarden A. 2017 Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat. Methods* **14**, 935–936. (doi:10.1038/nmeth.4437)
36. Zheng GXY *et al.* 2017 Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049. (doi:10.1038/ncomms14049)
37. Hughes AJ, Spelke DP, Xu Z, Kang C-C, Schaffer DV, Herr AE. 2014 Single-cell western blotting. *Nat. Methods* **11**, 749–755. (doi:10.1038/nmeth.2992)
38. Shi Q *et al.* 2012 Single-cell proteomic chip for profiling intracellular signaling pathways in single tumor cells. *Proc. Natl Acad. Sci. USA* **109**, 419–424. (doi:10.1073/pnas.1110865109)
39. Helmdach L, Schwartz F, Ulrich J. 2013 Process control using advanced particle analyzing systems: applications from crystallization to fermentation processes. *Chem. Eng. Technol.* **37**, 213–220. (doi:10.1002/ceat.201300190)
40. Brognaux A, Bugge J, Schwartz FH, Thonart P, Telek S, Delvigne F. 2013 Real-time monitoring of cell viability and cell density on the basis of a three dimensional optical reflectance method (3D-ORM): investigation of the effect of sub-lethal and lethal injuries. *J. Industr. Microbiol. Biotechnol.* **40**, 679–686. (doi:10.1007/s10295-013-1271-9)
41. Cohen AA *et al.* 2008 Dynamic proteomics of individual cancer cells in response to a drug. *Science* **322**, 1511–1516. (doi:10.1126/science.1160165)
42. Feigelman J, Ganscha S, Hastreiter S, Schwarzfischer M, Filipczyk A, Schroeder T, Theis FJ, Marr C, Claassen M. 2016 Analysis of cell lineage trees by exact Bayesian inference identifies negative autoregulation of Nanog in mouse embryonic stem cells. *Cell Syst.* **3**, 480–490.e13. (doi:10.1016/j.cels.2016.11.001)
43. Kuzmanovska I, Milias-Argeitis A, Mikelson J, Zechner C, Khammash M. 2017 Parameter inference for stochastic single-cell dynamics from lineage tree data. *BMC Syst. Biol.* **11**, 1–13. (doi:10.1186/s12918-017-0425-1)
44. Munsky B, Khammash M. 2006 The finite state projection algorithm for the solution of the chemical master equation. *J. Chem. Phys.* **124**, 044104. (doi:10.1063/1.2145882)
45. Waldherr S, Haasdonk B. 2012 Efficient parametric analysis of the chemical master equation through model order reduction. *BMC Syst. Biol.* **6**, 81. (doi:10.1186/1752-0509-6-81)
46. Gillespie CS. 2009 Moment-closure approximations for mass-action models. *IET Syst. Biol.* **3**, 52–58. (doi:10.1049/iet-syb:20070031)
47. Singh A, Hespanha JP. 2011 Approximate moment dynamics for chemically reacting systems. *IEEE Trans. Autom. Control* **56**, 414–418. (doi:10.1109/TAC.2010.2088631)
48. Schnoerr D, Sanguinetti G, Grima R. 2017 Approximation and inference methods for stochastic biochemical kinetics—a tutorial review. *J. Phys. A: Math. Theor.* **50**, 093001. (doi:10.1088/1751-8121/aa54d9)
49. Gillespie DT. 2000 The chemical Langevin equation. *J. Chem. Phys.* **113**, 297–306. (doi:10.1063/1.481811)
50. Munsky B, Trinh B, Khammash M. 2009 Listening to the noise: random fluctuations reveal gene network parameters. *Mol. Syst. Biol.* **5**, 1–7. (doi:10.1038/msb.2009.75)
51. Swain PS, Elowitz MB, Siggia ED. 2002 Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc. Natl Acad. Sci. USA* **99**, 12 795–12 800. (doi:10.1073/pnas.162041399)
52. Zechner C, Ruess J, Krenn P, Pelet S, Peter M, Lygeros J, Koeppl H. 2012 Moment-based inference predicts bimodality in transient gene expression. *Proc. Natl Acad. Sci. USA* **109**, 8340–8345. (doi:10.1073/pnas.1200161109)
53. Hasenauer J, Waldherr S, Doszczak M, Scheurich P, Radde N, Allgöwer F. 2011 Analysis of heterogeneous cell populations: a density-based modeling and identification framework. *J. Process Control* **21**, 1417–1425. (doi:10.1016/j.jprocont.2011.06.020)
54. Mantzaris NV, Daoutidis P, Srienc F. 2001 Numerical solution of multi-variable cell population balance models: I. Finite difference methods. *Comput. Chem. Eng.* **25**, 1411–1440. (doi:10.1016/S0098-1354(01)00709-8)
55. Mantzaris NV, Daoutidis P, Srienc F. 2001 Numerical solution of multi-variable cell population balance models. III. Finite element methods. *Comput. Chem. Eng.* **25**, 1463–1481. (doi:10.1016/S0098-1354(01)00711-6)
56. Kumar J, Peglow M, Warnecke G, Heinrich S, Mörl L. 2006 Improved accuracy and convergence of discretized population balance for aggregation: the cell average technique. *Chem. Eng. Sci.* **61**, 3327–3342. (doi:10.1016/j.ces.2005.12.014)
57. Mantzaris NV, Daoutidis P, Srienc F. 2001 Numerical solution of multi-variable cell population balance models. II. Spectral methods. *Comput. Chem. Eng.* **25**, 1441–1462. (doi:10.1016/S0098-1354(01)00710-4)
58. Bück A, Klaunick G, Kumar J, Peglow M, Tsotsas E. 2012 Numerical simulation of particulate processes for control and estimation by spectral methods. *AIChE J.* **58**, 2309–2319. (doi:10.1002/aic.12757)
59. Pinto MA, Immanuel CD, Doyle FJ III. 2007 A feasible solution technique for higher-dimensional population balance models. *Comput. Chem. Eng.* **31**, 1242–1256. (doi:10.1016/j.compchemeng.2006.10.016)
60. Ramkrishna D. 1979 Statistical models of cell populations. In *Advances in biochemical engineering*, vol. 11 (eds TK Ghose, A Fiechter, N Blakebrough), pp. 1–47. Berlin, Germany: Springer. (doi:10.1007/3-540-08990-X_21)
61. Liou JJ, Srienc F, Fredrickson AG. 1997 Solutions of population balance models based on a successive generations approach. *Chem. Eng. Sci.* **52**, 1529–1540. (doi:10.1016/s0009-2509(96)00510-6)
62. Mantzaris NV. 2006 Stochastic and deterministic simulations of heterogeneous cell population dynamics. *J. Theor. Biol.* **241**, 690–706. (doi:10.1016/j.jtbi.2006.01.005)
63. Mantzaris NV. 2007 From single-cell genetic architecture to cell population dynamics: quantitatively decomposing the effects of different population heterogeneity sources for a genetic network with positive feedback architecture. *Biophys. J.* **92**, 4271–4288. (doi:10.1529/biophysj.106.100271)
64. Silverman BW. 1986 *Density estimation for statistics and data analysis*. Boca Raton, FL: CRC Press.
65. Stamatakis M. 2013 Cell population balance and hybrid modeling of population dynamics for a single gene with feedback. *Comput. Chem. Eng.* **53**, 25–34. (doi:10.1016/j.compchemeng.2013.02.006)
66. Waldherr S, Trennt P, Hussain M. 2016 Hybrid simulation of heterogeneous cell populations. *IEEE Life Sci. Lett.* **2**, 9–12. (doi:10.1109/LLS.2016.2615089)
67. Zeng S, Waldherr S, Ebenbauer C, Allgöwer F. 2016 Ensemble observability of linear systems. *IEEE Trans. Autom. Control* **61**, 1452–1465. (doi:10.1109/TAC.2015.2463631)
68. Waldherr S, Frysch R, Pfeiffer T, Jakuszeit T, Zeng S, Rose G. 2015 A numerical evaluation of state reconstruction methods for heterogeneous cell

populations. In *Proc. European Control Conf. (ECC), Linz, Austria, 15–17 July 2015*, pp. 2926–2931.

69. Spetsieris K, Zygourakis K, Mantzaris NV. 2009 A novel assay based on fluorescence microscopy and image processing for determining phenotypic distributions of rod-shaped bacteria. *Biotechnol. Bioeng.* **102**, 598–615. (doi:10.1002/bit.22063)

70. Spetsieris K, Zygourakis K. 2012 Single-cell behavior and population heterogeneity: solving an inverse problem to compute the intrinsic physiological state functions. *J. Biotechnol.* **158**, 80–90. (doi:10.1016/j.jbiotec.2011.08.018)

71. Hussain M, Waldherr S. 2016 Extraction of physiological state functions in heterogeneous cell population models. In *Proc. 6th IFAC Symp. on Foundations of Systems Biology in Engineering, Magdeburg, Germany, 9–12 October 2016*, pp. 258–263.

72. Loo L-H, Lin H-J, Singh DK, Lyons KM, Altschuler SJ, Wu LF. 2009 Heterogeneity in the physiological states and pharmacological responses of differentiating 3T3-L1 preadipocytes. *J. Cell Biol.* **187**, 375–384. (doi:10.1083/jcb.200904140)

73. Isensee J, Diskar M, Waldherr S, Buschow R, Hasenauer J, Prinz A, Allgöwer F, Herberg FW, Hucho T. 2014 Pain modulators regulate the dynamics of PKA-RII phosphorylation in subgroups of sensory neurons. *J. Cell Sci.* **127**, 216–229. (doi:10.1242/jcs.136580)

74. Loos C, Moeller K, Fröhlich F, Hucho T, Hasenauer J. 2018 Mechanistic hierarchical population model identifies latent causes of cell-to-cell variability. *Cell Syst.* **6**, 593–603.e13. (doi:10.1016/j.cels.2018.04.008)

75. Perlman ZE, Slack MD, Feng Y, Mitchison TJ, Wu LF, Altschuler SJ. 2004 Multidimensional drug profiling by automated microscopy. *Science* **306**, 1194–1198. (doi:10.1126/science.1100709)

76. Shu C-C, Chatterjee A, Dunny G, Hu W-S, Ramkrishna D. 2011 Bistability versus bimodal distributions in gene regulatory processes from population balance. *PLoS Comput. Biol.* **7**, e1002140. (doi:10.1371/journal.pcbi.1002140)

77. Hawkins ED, Turner ML, Dowling MR, van Gend C, Hodgkin PD. 2007 A model of immune regulation as a consequence of randomized lymphocyte division and death times. *Proc. Natl Acad. Sci. USA* **104**, 5032–5037. (doi:10.1073/pnas.0700026104)

78. Glauche I, Moore K, Thielecke L, Horn K, Loeffler M, Roeder I. 2009 Stem cell proliferation and quiescence—two sides of the same coin. *PLoS Comput. Biol.* **5**, e1000447. (doi:10.1371/journal.pcbi.1000447)

79. Johnston MD, Edwards CM, Bodmer WF, Maini PK, Jonathan Chapman S. 2007 Mathematical modeling of cell population dynamics in the colonic crypt and in colorectal cancer. *Proc. Natl Acad. Sci. USA* **104**, 4008–4013. (doi:10.1073/pnas.0611179104)

80. Doumic M, Marciniak-Czochra A, Perthame B, Zubelli JP. 2011 A structured population model of cell differentiation. *SIAM J. Appl. Math.* **71**, 1918–1940. (doi:10.1137/100816584)

81. Waldherr S, Wu J, Allgöwer F. 2010 Bridging time scales in cellular decision making with a stochastic bistable switch. *BMC Syst. Biol.* **4**, 108–112. (doi:10.1186/1752-0509-4-108)

82. Ahmed R *et al.* 2015 Reconciling estimates of cell proliferation from stable isotope labeling experiments. *PLoS Comput. Biol.* **11**, e1004355. (doi:10.1371/journal.pcbi.1004355)

83. Schittler D, Allgöwer F, DeBoer RJ. 2013 A new model to simulate and analyze proliferating cell populations in BrdU labeling experiments. *BMC Syst. Biol.* **7**, S4. (doi:10.1186/1752-0509-7-S1-S4)

84. Hasenauer J, Schittler D, Allgöwer F. 2012 Analysis and simulation of division- and label-structured population models. *Bull. Math. Biol.* **8**, 227–241. (doi:10.1007/s11538-012-9774-5)

85. Perfahl H *et al.* 2011 Multiscale modelling of vascular tumour growth in 3D: the roles of domain size and boundary conditions. *PLoS ONE* **6**, e14790. (doi:10.1371/journal.pone.0014790)

86. Wang Z, Butner JD, Kerketta R, Cristini V, Deisboeck TS. 2015 Simulating cancer growth with multiscale agent-based modeling. *Semin. Cancer Biol.* **30**, 70–78. (doi:10.1016/j.semcancer.2014.04.001)

87. Henson MA, Müller D, Reuss M. 2002 Cell population modelling of yeast glycolytic oscillations. *Biochem. J.* **368**, 433–446. (doi:10.1042/bj20021051)

88. To T-L, Henson MA, Herzog ED, Doyle FJ III. 2007 A molecular model for intercellular synchronization in the mammalian circadian clock. *Biophys. J.* **92**, 3792–3803. (doi:10.1529/biophysj.106.094086)

89. Lang M, Marquez-Lago T, Stelling J, Waldherr S. 2011 Autonomous synchronization of chemically coupled synthetic oscillators. *Bull. Math. Biol.* **73**, 2678–2706. (doi:10.1007/s11538-011-9642-8)

90. Olav Hald B, Garkier Hendriksen M, Graae Sørensen P. 2013 Programming strategy for efficient modeling of dynamics in a population of heterogeneous cells. *Bioinformatics* **29**, 1292–1298. (doi:10.1093/bioinformatics/btt132)

91. Wolf J, Heinrich R. 2000 Effect of cellular interaction on glycolytic oscillations in yeast: a theoretical investigation. *Biochem. J.* **345**, 321–334. (doi:10.1042/bj3450321)

92. Glauche I, Lorenz R, Hasenclever D, Roeder I. 2009 A novel view on stem cell development: analysing the shape of cellular genealogies. *Cell Prolif.* **42**, 248–263. (doi:10.1111/j.1365-2184.2009.00586.x)

93. Gonzalez-Vargas AM, Cinquemani E, Ferrari-Trecate G. 2016 Validation methods for population models of gene expression dynamics. *IFAC-PapersOnLine* **49**, 114–119. (doi:10.1016/j.ifacol.2016.12.112)

94. Gillespie DT. 1977 Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81**, 2340–2361. (doi:10.1021/j100540a008)

95. Wolkenhauer O. 2007 Defining systems biology: an engineering perspective. *IET Syst. Biol.* **1**, 204–206. (doi:10.1049/iet-syb:20079017)

96. Gauthier JP, Kupka IAK. 1994 Observability and observers for nonlinear systems. *SIAM J. Control Optimiz.* **32**, 975–994. (doi:10.1137/S0363012991221791)

97. Berthoumieux S, Brilli M, Kahn D, de Jong H, Cinquemani E. 2013 On the identifiability of metabolic network models. *J. Math. Biol.* **67**, 1795–1832. (doi:10.1007/s00285-012-0614-x)

98. Raue A, Kreutz C, Maiwald T, Bachmann J, Schilling M, Klingmüller U, Timmer J. 2009 Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics* **25**, 1923–1929. (doi:10.1093/bioinformatics/btp358)

99. Waldherr S, Zeng S, Allgöwer F. 2014 Identifiability of population models via a measure theoretical approach. In *Proc. 19th World Congress of the International Federation of Automatic Control (IFAC), Cape Town, South Africa, 24–29 August 2014*, pp. 1717–1722.

100. Zeng S, Waldherr S, Allgöwer F. 2014 An inverse problem of tomographic type in population dynamics. In *Proc. 53rd IEEE Conf. on Decision and Control (CDC), Los Angeles, CA, USA, 15–17 December 2014*, pp. 1643–1648. (doi:10.1109/CDC.2014.7039635)

101. Küper A, Dürr R, Waldherr S. In press. Dynamic density estimation in heterogeneous cell populations. *IEEE Control Syst. Lett.* (doi:10.1109/lcsys.2018.2847905)

102. Waldherr S, Hasenauer J, Allgöwer F. 2009 Estimation of biochemical network parameter distributions in cell populations. In *Proc. 15th IFAC Symp. on System Identification, Saint-Malo, France, 6–8 July 2009*, pp. 1265–1270.

103. Dixit P, Lyashenko E, Niepel M, Vitkup D. 2018 Maximum entropy framework for inference of cell population heterogeneity in signaling network dynamics. *bioRxiv*. (doi:10.1101/137513)

104. Landweber L. 1951 An iteration formula for Fredholm integral equations of the first kind. *Am. J. Math.* **73**, 615–624. (doi:10.2307/2372313)

105. Raue A *et al.* 2013 Lessons learned from quantitative dynamical modeling in systems biology. *PLoS ONE* **8**, e74335. (doi:10.1371/journal.pone.0074335)

106. Moles CG, Mendes P, Banga JR. 2003 Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome Res.* **13**, 2467–2474. (doi:10.1101/gr.1262503)

107. Balsa-Canto E, Peifer M, Banga JR, Timmer J, Fleck C. 2008 Hybrid optimization method with general switching strategy for parameter estimation. *BMC Syst. Biol.* **2**, 26. (doi:10.1186/1752-0509-2-26)

108. Banks HT, Kenz ZR, Thompson WC. 2012 A review of selected techniques in inverse problem nonparametric probability distribution estimation. *J. Inv. Ill-posed Probl.* **20**, 429–460. (doi:10.1515/jip-2012-0037)

109. Collins JF, Richmond MH. 1962 Rate of growth of *Bacillus cereus* between divisions. *J. Gen. Microbiol.* **28**, 15–33. (doi:10.1099/00221287-28-1-15)

110. Sathyagal AN, Ramkrishna D, Narsimhan G. 1995 Solution of inverse problems in population

balances-II. Particle break-up. *Comput. Chem. Eng.* **19**, 437–451. (doi:10.1016/0098-1354(94)00062-S)

111. Mahoney AW, Doyle FJ III, Ramkrishna D. 2002 Inverse problems in population balances: growth and nucleation from dynamic data. *AIChE J.* **48**, 981–990. (doi:10.1002/aic.690480508)
112. Groh A, Kohr H, Louis AK. 2016 Numerical rate function determination in partial differential equations modeling cell population dynamics. *J. Math. Biol.* **74**, 533–565. (doi:10.1007/s00285-016-1032-2)
113. Veiga-Fernandes H, Walter U, Bourgeois C, McLean A, Rocha B. 2000 Response of naïve and memory CD8$^+$ T cells to antigen stimulation *in vivo*. *Nat. Immunol.* **1**, 47–53. (doi:10.1038/76907)
114. Revy P, Sospedra M, Barbour B, Trautmann A. 2001 Functional antigen-independent synapses formed between T cells and dendritic cells. *Nat. Immunol.* **2**, 925–931. (doi:10.1038/ni713)
115. De Boer RJ, Perelson AS. 2005 Estimating division and death rates from CFSE data. *J. Comput. Appl. Math.* **184**, 140–164. (doi:10.1016/j.cam.2004.08.020)
116. Luzyanina T, Roose D, Schenkel T, Sester M, Ehl S, Meyerhans A, Bocharov G. 2007 Numerical modelling of label-structured cell population growth using CFSE distribution data. *Theor. Biol. Med. Model.* **4**, 26. (doi:10.1186/1742-4682-4-26)
117. Luzyanina T, Roose D, Bocharov G. 2008 Distributed parameter identification for a label-structured cell population dynamics model using CFSE histogram time-series data. *J. Math. Biol.* **59**, 581–603. (doi:10.1007/s00285-008-0244-5)
118. Banks HT, Sutton KL, Clayton Thompson W, Bocharov G, Roose D, Schenkel T, Meyerhans A. 2010 Estimation of cell proliferation dynamics using CFSE data. *Bull. Math. Biol.* **73**, 116–150. (doi:10.1007/s11538-010-9524-5)
119. Hasenauer J. 2012 Modeling and parameter estimation for heterogeneous cell populations. PhD thesis, University of Stuttgart, Germany.
120. Hross S, Hasenauer J. 2016 Analysis of CFSE time-series data using division-, age- and label-structured population models. *Bioinformatics* **32**, 2321–2329. (doi:10.1093/bioinformatics/btw131)