Estimation of biochemical network parameter distributions in cell populations

Steffen Waldherr [1,2], **Jan Hasenauer** [2], and **Frank Allgöwer**

Institute for Systems Theory and Automatic Control,
Universität Stuttgart, Pfaffenwaldring 9, Stuttgart, Germany

Abstract: Populations of heterogeneous cells play an important role in many biological systems. In this paper we consider systems where each cell can be modelled by an ordinary differential equation. To account for heterogeneity, parameter values are different among individual cells, subject to a distribution function which is part of the model specification.

Experimental data for heterogeneous cell populations can be obtained from flow cytometric fluorescence microscopy. We present a heuristic approach to use such data for estimation of the parameter distribution in the population. The approach is based on generating simulation data for samples in parameter space. By convex optimisation, a suitable probability density function for these samples is computed.

To evaluate the proposed approach, we consider artificial data from a simple model of the tumor necrosis factor (TNF) signalling pathway. Its main characteristic is a bimodality in the TNF response: a certain percentage of cells undergoes apoptosis upon stimulation, while the remaining part stays alive. We show how our modelling approach allows to identify the reasons that underly the differential response. *Copyright © 2009 IFAC.*

Keywords: estimation of probability distribution, population models, TNF signalling

1. INTRODUCTION

Modelling in systems biology typically aims at achieving a quantitative description of intracellular signal transduction or differentiation processes at the cellular level. Most models describe a "typical single cell" on the basis of experimental data obtained from cell populations. However, to understand dynamical behaviour within heterogeneous cell populations, a consideration of many cells within the whole population is mandatory.

Phenotypic heterogeneity in genetically identical cells arises mainly from stochasticity in biochemical reactions, unequal partitioning of cellular material at cell division (Mantzaris, 2007), or epigenetic differences (Avery, 2006). When considering cells with high mutation rate, such as cancer cells, also genotypic heterogeneity plays a major role. For this paper, we choose to model heterogeneity by differences in parameter values of the model describing the process of interest. The model structure is on the contrary assumed to be identical in all cells, as it usually represents the physical interactions among molecules, which should be independent of the cell's state. The parametric approach is well suited for genetic and epigenetic differences. We assume that interactions among cells in the population can be neglected for the process to be studied. This is indeed the case for many relevant signalling pathways, and is also an implicit assumption in many single-cell models. The distribution of parameter values within the considered cell population is described by a suitable multivariate probability distribution function, which needs to

be part of the model specification. Mathematical modelling of such a process will typically result in a non-linear partial differential equation for the probability distributions of the state variables (Mantzaris, 2007). Since this is very hard to deal with, we propose to use a sample based approach, consisting of a large collection of ordinary differential equation systems of identical structure, but with differing parameter values which are subject to a specified parameter distribution function.

In this paper, we explore the possibility to estimate the parameter distribution function using experimental large-scale measurements of the distributions of system variables within the cell population. Such data is available on a suitable scale from a newly developed measurement technology, the flow cytometric fluorescence microscopy (Ortyn et al., 2007), which is a combination of classical flow cytometry and fluorescence microscopy.

Classical flow cytometry is a long-established tool to obtain distributions of system variables in heterogeneous cell populations (Perez and Nolan, 2006). To measure the activity of signalling proteins, suitable fluorescence markers are introduced into the cells. A stream of several thousand cells per second is then injected into the measurement device, and the fluorescence intensity of each individual cell can be measured. While static flow cytometry measurements are very common in experimental setups, corresponding time course data are rarely collected and are typically quite sparse (Gardner et al., 2000).

Fluorescence microscopy is another established experimental tool, where microscopic images from a population of fluorescently labelled cells are collected and evaluated by

---

[1] Corresponding author (`waldherr@ist.uni-stuttgart.de`)
[2] Both authors contributed equally to this work.

image analysis. With the classical technical implementation, fluorescence microscopy is limited to small sample numbers. Yet percentages of cells showing a particular feature, such as an apoptotic phenotype, are commonly measured at several time instances. Also distributions of relevant variables over a time course have been measured (Mettetal et al., 2006), but the technology is not widely used in dynamical modelling studies.

Flow cytometric fluorescence microscopy now combines flow cytometry with single cell fluorescence microscopy by taking microscopic images of individual cells while they pass the flow cytometer. This allows to collect and analyse microscopic images of several thousand fluorescently labelled cells per minute (Ortyn et al., 2007), with technological requirements similar to classical flow cytometry. In this way, distributions of signalling protein activities can be measured efficiently in large populations of heterogeneous cells (George et al., 2006). Although the technology has not been used so far to obtain distributions of relevant variables at several time instances, such measurements are now becoming experimentally feasible.

Estimation of parameter distributions in model collections that represent a heterogeneous population is a long-standing topic in pharmacodynamics (Al-Banna et al., 1990). However, a crucial difference between pharmacological experiments and cell population measurements is that in pharmacodynamics, samples are taken from the same individuals at all time points, measurements are linked to individuals, and as a consequence individual trajectories are known. This is not the case in fluorescence microscopy, where each individual cell is measured only once, and for each time point only the distribution of the measured variable within the population is recorded.

Other established approaches to parameter estimation of probabilistic systems usually consider a problem setup where the output of a single cell is directly considered as available measurement data at all time instances. This is quite different to our setup, where each individual cell can only be taken for measurement once, and thus only the distribution of output variables within the population is reliably known for all sampling times. As a consequence, established approaches of parameter estimation seem not to be well suited to deal with this problem.

In this paper, we present a heuristic approach to estimate the parameter distribution from the distributions of measured variables. In a first step, simulation data is generated for a suitable choice of parameter samples. As such, the approach is related to classical particle filters (Doucet et al., 2001). However, instead of an iterative updating, we construct a convex optimisation problem that produces a suitable weighting for the considered parameter samples. This weighting can directly be transformed into a probability distribution for the parameter values.

The paper is structured as follows. In Section 2, the population modelling framework that we are using is introduced. In Section 3, we present the proposed method to estimate the parameter distribution function of the model based on population measurements. Section 4 describes the application of the proposed method to simulated data for a model of the TNF signalling pathway, and discusses how to use population modelling in order to evaluate differences in cellular behavior within a heterogeneous cell population.

Notation: Denote by $[\check{z}, \hat{z}] \subset \mathbb{R}^k$ the hyperrectangle $\{z \in \mathbb{R}^k : \check{z}_i \leq z_i \leq \hat{z}_i, i = 1, \dots, k\}$.

2. PARAMETER-DISTRIBUTED POPULATION MODELS

For the purpose of this paper, a model of a biochemical reaction network in a population of N cells is given by the collection of differential equations

$$\begin{aligned} \dot{x}^{(i)}(t) &= f(x^{(i)}(t), \pi^{(i)}), \quad x^{(i)}(0) = x_0^{(i)}, \\ y^{(i)}(t) &= h(x^{(i)}(t)), \qquad\qquad i = 1, \dots, N \end{aligned} \quad (1)$$

with state variables $x(t) \in \mathbb{R}^n$, measured variables $y(t) \in \mathbb{R}^q$, and parameters $\pi \in \mathbb{R}^r$. The index i specifies the individual cells within the population. We collect the parameters and initial condition in the extended parameter vector $p^{(i)} = (\pi^{(i)}, x_0^{(i)}) \in \mathbb{R}^m$, where $m = n + r$.

We assume that the population is heterogeneous, where heterogeneity is accounted for by differences in parameter values among individual cells. The distribution of parameters and initial conditions is given by a cumulative probability distribution function $\Phi : \mathbb{R}^m \to [0, 1]$ which is part of the model specification, i.e. parameter values and initial conditions for the cell with index i are subject to the probability distribution

$$\text{Prob}(p_1^{(i)} \leq p_1, \dots, p_m^{(i)} \leq p_m) = \Phi(p_1, \dots, p_m). \quad (2)$$

Due to the measurement technology, the output of every individual cell can only be measured once during the course of an experiment, because the cell is removed from the population for the measurement. Thus, instead of considering the measured output $y^{(i)}$ directly, it makes more sense to consider the distribution of $y^{(i)}$ at sampling times t_k, $k = 1, \dots, K$ as an output. At each sampling time, M cells are selected arbitrarily from the population and subjected to measurement. We assume that M is large enough such that a reliable approximation of the output distribution within the whole population can be obtained. The measurement taken from (1) is thus given by functions

$$\Psi_k(y) = \text{Prob}(y^{(i)}(t_k) \leq y), \quad k = 1, \dots, K, \ i = 1, \dots, N. \quad (3)$$

The goal of parameter distribution estimation is to compute the function $\Phi(p_1, \dots, p_m)$ from knowledge of the functions $\Psi_k(y)$ and the model structure (1). Typically the measurement of $\Psi_k(y)$ is discretized over suitable hyperrectangles in the variable y. Since the number of cells being measured is finite, the values of $\Psi_k(y)$ are discrete as well, although for most considerations we can assume the number of cells large enough to neglect this.

3. PARAMETER ESTIMATION METHOD FOR POPULATION MODELS

3.1 Maximum Entropy approach to probability density estimation

The proposed estimation method is based on the simulation of (1) for all parameter values and initial conditions contained in a finite sample

$$\mathcal{P} = \{p^{(i)} \in \mathbb{R}^m : i = 1, \dots, M\}.$$

A good choice for a sampling set is the so called Latin hypercube, which ensures that the total range of the relevant parameter set is captured (Stein, 1987).

Definition 1. A finite set $\mathcal{P} \subset [\check{p}, \hat{p}] \subset \mathbb{R}^m$ is called a *latin hypercube in $\mathbb{R}^m$ with sampling density d*, if it contains exactly one point $p \in \mathbb{R}^m$ such that

$$(\alpha - 1)\frac{\hat{p}_i - \check{p}_i}{d} < p_i - \check{p}_i < \alpha \frac{\hat{p}_i - \check{p}_i}{d} \qquad (4)$$

for each $i = 1, \ldots, m$ and $\alpha = 1, \ldots, d$.

Having chosen the sampling set $\mathcal{P}$ as a Latin hypercube, the goal is to estimate the fraction of cells, $\varphi(p^{(i)})$, which have an extended parameter vector close to $p^{(i)}$, such that the weighted simulated trajectories approximate the measured population dynamics reasonably well. This fraction is a measure for the relative contribution of the neighbourhood of $p^{(i)}$ to the cell population response. Additionally, the fractions have to sum up to one,

$$\sum_{i=1}^{M} \varphi(p^{(i)}) = 1. \qquad (5)$$

Hence, $\varphi(p^{(i)})$ can be interpreted as an approximation of the probability density function at $p^{(i)}$.

In order to calculate $\varphi(p^{(i)})$, the output space is divided into hyperrectangles. For each sampling time, the fraction of cells of the population which is contained in each hyperrectangle is computed according to the following definition. An illustration is shown in Figure 1.

Definition 2. A q-dimensional array $Y(t_k) \in \mathbb{R}^{\beta_1 \times \ldots \times \beta_q}$ is called a *discretized population distribution* at time t_k with discretization vector $\beta = [\beta_1, \ldots, \beta_q]$, if

$$Y_{(\gamma_1, \ldots, \gamma_q)}(t_k) = \Psi_k(\hat{y}^\gamma) - \Psi_k(\check{y}^\gamma), \qquad (6)$$

where $\gamma_1 = 1, \ldots, \beta_1, \ldots, \gamma_q = 1, \ldots, \beta_q$, and $\check{y}_i^\gamma = \check{y}_i + (\gamma_i - 1)\frac{\hat{y}_i - \check{y}_i}{\beta_i}$ and $\hat{y}_i^\gamma = \check{y}_i + \gamma_i \frac{\hat{y}_i - \check{y}_i}{\beta_i}$, $i = 1, \ldots, m$. Hereby $\check{y}_i$ and $\hat{y}_i$ are the minimal respectively maximal values of output i which are measured.

The array $Y(t_k)$ can be interpreted as a discrete approximation of the probability density function of the outputs. To compute the $\varphi(p^{(i)})$, $i = 1, \ldots, M$, the system (1) is simulated for every $p^{(i)} \in \mathcal{P}$ and the obtained outputs are discretized.

Definition 3. A q-dimensional array $\tilde{Y}^{(i)}(t_k) \in \mathbb{R}^{\beta_1 \times \ldots \times \beta_q}$ is called a *discretized trajectory of $\tilde{y}^{(i)}$* at time points t_k, $k = 1, \ldots, K$, with discretization vector $\beta = [\beta_1, \ldots, \beta_q]$, if
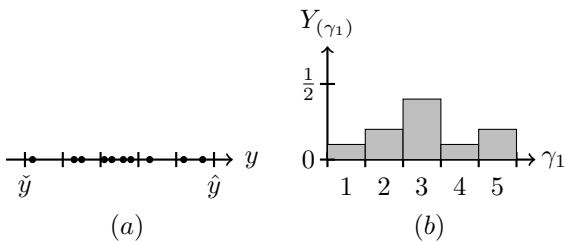

Fig. 1. Illustration of the discretized population distribution. (a) shows as dots the measured outputs at time t_k and (b) depicts the corresponding discretized population distribution for $\beta = 5$.

$$\tilde{Y}^{(i)}_{(\gamma_1, \ldots, \gamma_q)}(t_k) = \begin{cases} 1, & \text{if } \tilde{y}^{(i)}(t_k) \in [\check{y}^\gamma, \hat{y}^\gamma] \\ 0, & \text{otherwise}, \end{cases} \qquad (7)$$
$$\gamma_1 = 1, \ldots, \beta_1, \ldots, \gamma_q = 1, \ldots, \beta_q.$$

Hereby $\tilde{y}^{(i)}$ is the output of the system obtained by simulation with $p^{(i)}$. $\check{y}^\gamma$, $\hat{y}^\gamma$ are defined as before.

The array $\tilde{Y}^{(i)}(t_k)$ can also be interpreted as an approximation of the probability density function for cells with the parameter vector $p^{(i)}$. Given the discretized measured population dynamics and the discretized simulated trajectories, we intend to compute the fractions, $\varphi(p^{(i)})$, such that the difference between the weighted sum of simulated trajectories and the measured population dynamics,

$$\Delta Y(\varphi, t_k) = \sum_{i=1}^{M} \varphi(p^{(i)}) \tilde{Y}^{(i)}(t_k) - Y(t_k), \qquad (8)$$

is zero, for $k = 1, \ldots, K$, where

$$\varphi = \left[\varphi(p^{(1)}), \ldots, \varphi(p^{(M)})\right]^T. \qquad (9)$$

This problem could be solved using standard least square techniques, but this leads in many cases to dramatical overfitting. Especially, if the measurement data does not contain enough information to fit the parameter distribution, a spiky probability density function is obtained. Least square techniques generally select a minimum norm solution for underdetermined systems, so it can happen that parameters which are not identifiable show a peak in the probability density function at a single point.

This should be avoided, because it is desirable that the resulting distribution indicates whether a parameter is identifiable given the measured data or not. Therefore, we choose an entropy based approach to determine $\varphi(p^{(i)})$.

Definition 4. The function $\text{Ent} : \mathbb{R}^M \to \mathbb{R}$ given by

$$\text{Ent}(\varphi) = -\sum_{i=1}^{M} \varphi(p^{(i)}) \ln(\varphi(p^{(i)})). \qquad (10)$$

is called the *entropy* of φ.

Given an underdetermined system of equations, the maximum entropy approach selects the solution which contains least information, and thus avoids adding artificial information to the measurement data (MacKay, 2003). In our case this implies that the "flattest" probability distribution which fits all the constraints is selected in the optimisation problem. Thus, if a parameter is not identifiable, we obtain a very flat distribution and no information is added.

The entropy approach yields the optimisation problem

$$\begin{array}{ll} \max & \text{Ent}(\varphi) \\ \text{s.t.} & \Delta Y(\varphi, t_k) = 0, \ k = 1, \ldots, K \\ & \mathbf{1}^T \varphi = 1 \\ & \varphi \geq 0, \end{array} \qquad (11)$$

where $\mathbf{1} = [1, \ldots, 1]^T \in \mathbb{R}^M$. The solution of (11) is the weighting vector φ, with the highest entropy which exactly reproduces the discretized measured population dynamics. Unfortunately, (11) is very likely to be infeasible because even if the equation

$$\Delta Y(\varphi, t_k) = 0, \quad k = 1, \ldots, K, \qquad (12)$$

is underdetermined, it cannot be ensured that a solution exists. Reasons are measurement errors and small cell

numbers in measurements, but primarily an insufficient discretization of the parameter space. To improve the feasibility, small discrepancies between the measured and the weighted simulated population are allowed. This leads to the relaxed problem

$$\begin{aligned} \max \quad & \text{Ent}(\varphi) \\ \text{s.t.} \quad & \Delta Y(\varphi, t_k) \in [-\hat{e}, \hat{e}], \ k = 1, \dots, K \\ & \mathbf{1}^T \varphi = 1 \\ & \varphi \geq 0, \end{aligned} \tag{13}$$

where $\Delta Y(\varphi, t_k) \in [-\hat{e}, \hat{e}]$ denotes the constraint that each element of ΔY is bounded between $-\hat{e}$ and $+\hat{e}$. As before, the other constraints are that all weights sum up to one, and that all weights are greater than or equal to zero.

We are left with the problem to define the error bound $\hat{e}$. A known constraint is that $\hat{e} \in [0, 1]$. To obtain estimation results that fit the measurements as good as possible, $\hat{e}$ is decreased to the minimal value for which (13) is still feasible. This is done via a bisection algorithm.

The relaxed optimisation problem (13) is convex. The entropy is concave and the constraints are linear. For the class of convex optimisation problems efficient solvers exist, for instance the primal-dual-interior point method (Boyd and Vandenberghe, 2004). Convergence to the global maximum in polynomial time can be guaranteed.

Based on the solution of (13), an estimate $\hat{\Phi}$ for the parameter distribution function Φ is computed as

$$\hat{\Phi}(p) = \sum_{i : p \geq p^{(i)} \in \mathcal{P}} \varphi(p^{(i)}). \tag{14}$$

3.2 Distribution estimation for independent parameters

A simplifying yet convenient assumption is that parameter values and initial conditions are independently distributed. Although not strictly true in most cases, it is reasonable to make this simplification also if parameters are only weakly correlated. In this case, the probability distribution function can be decomposed as

$$\Phi(p) = \Phi_1(p_1)\Phi_2(p_2) \cdots \Phi_m(p_m), \tag{15}$$

where $\Phi_i(p_i)$ denotes the distribution function for the i-th parameter.

Based on the estimate $\hat{\Phi}$ (14), estimates for the individual distribution functions can be computed by marginalising the other parameters, i.e. taking $\Phi_i(p_i) = \lim_{p_j \to \infty, j \neq i} \Phi(p)$. Thus an estimate $\hat{\Phi}_i$ for the individual distributions is obtained as

$$\hat{\Phi}_i(p_i) = \sum_{j : p_i \geq p_i^{(j)} \in \mathcal{P}} \varphi(p^{(j)}) \tag{16}$$

for $i = 1, \dots, m$.

4. APPLICATION TO A TNF SIGNAL TRANSDUCTION MODEL

4.1 Motivation for population modelling

TNF is a signalling hormone involved in the inflammatory response of mammalian cells. It can induce programmed cell death (apoptosis) via the caspase cascade, but has also anti-apoptotic effects via the NF-κB pathway (Wajant

i	1	2	3	4	5
a_i	0.6	0.2	0.2	0.5	
b_i	0.4	0.7	0.3	0.5	0.4

Table 1. Nominal parameter values for the TNF signalling model (17).

et al., 2003). For many cell types, stimulation with TNF will incude apoptosis in a certain percentage of the population, while the remaining cells stay alive. The reasons for this heterogeneous behaviour are unclear, but of great interest for biological research in TNF signalling. However, a major obstacle to the direct experimental study of the process is that the behaviour of individual cells cannot be monitored on a population scale over the time scale of interest. To overcome this problem, we propose the use of population modelling and estimation of parameter distributions from experimental population data. With a suitable model, a collection of single cell trajectories can be clustered according to the individual cell's fate and compared for characteristic differences in parameters or early-stage cell behaviour.

In this paper, we use artificial measurement data, generated from simulations, for two reasons. First, suitable experimental data is not yet available. Second, the purpose of the paper is more an evaluation of the estimation method itself than its application in biological research. Also, since no general results on parameter identifiability in the considered problem are available, such a study should be done in each application of the method to evaluate identifiability properties.

4.2 Presentation of the TNF signal transduction model

The model is based on earlier work from Chaves et al. (2008) and is built from known activating and inhibitory interactions among key signalling proteins. It includes as state variables activities of the caspases 8 and 3 (C8, C3), the transcription factor NF-κB and its inhibitor I-κB. The model is given by the ODE system

$$\begin{aligned} \dot{x}_1 &= -x_1 + \frac{1}{2}(\beta_4(x_3)\alpha_1(u) + \alpha_3(x_2)) \\ \dot{x}_2 &= -x_2 + \alpha_2(x_1)\beta_3(x_3) \\ \dot{x}_3 &= -x_3 + \beta_2(x_2)\beta_5(x_4) \\ \dot{x}_4 &= -x_4 + \frac{1}{2}(\beta_1(u) + \alpha_4(x_3)). \end{aligned} \tag{17}$$

The state variables x_i, $i = 1, \dots, 4$ are bounded between 0 and 1 and denote the relative activities of the signalling proteins C8, C3, NF-κB and I-κB, respectively. The functions $\alpha_j(x_i)$, $j = 1, \dots, 4$ represent activating connections and are given by $\alpha_j(x_i) = \frac{x_i^2}{a_j^2 + x_i^2}$. Correspondingly, $\beta_j(x_i) = \frac{b_j^2}{b_j^2 + x_i^2}$, $j = 1, \dots, 5$ represent inhibiting connections. a_j and b_j are parameters with values between 0 and 1, representing activation and inhibition thresholds, respectively. The input u denotes the external TNF stimulus. Nominal parameter values are given in Table 1.

4.3 Results of parameter distribution estimation

To evaluate the proposed approach we consider a virtual experimental setup in which the caspase 3 and
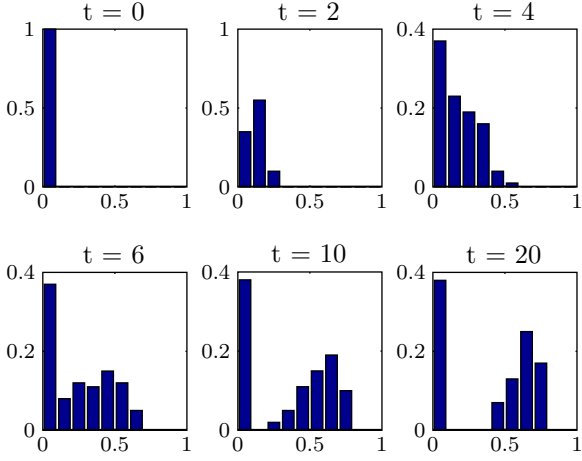
Fig. 2. Measured output distributions for caspase 3 activity
at the considered sampling times. The horizontal axis
gives the caspase 3 activity, the size of the bars
indicates the relative frequency in the population.

NF-κB activity is measured at the time points $t \in \{0, 0.5, 1, 2, 4, 6, 8, 10, 15, 20\}$ by flow cytometric microscopy. For each time point, the outputs of 10000 simulated cells are obtained, resulting in an output density distribution for each time point. Measurement errors are neglected in this example to make the interpretation of the results as simple as possible.

For this example, we assume heterogeneity in the parameters a_1, a_4, b_2, and b_3. For the generation of measurement data, each of the heterogeneous parameters is assumed to be distributed according to a log-normal distribution around the nominal values given in table 1. Each cell is assumed to have an initial condition which corresponds to the steady state with $x_1 = x_2 = 0$ for $u = 0$, where x_3 and x_4 depend on the individual parameter values. The considered heterogeneity is interesting, because it results in a bimodal response of the population to a TNF pulse applied during the time interval $0 < t < 2$. The output distributions in caspase 3 activity that would be measured in this case are shown in Figure 2. About 35 % of the population returns to zero caspase 3 activity, while the remaining cells show sustained caspase activity. The probability density functions of the parameter values are shown in Figure 3.

For the identification of the parameter distribution, the lower and upper bounds of all parameters are set to be 0 respectively 1. The sampling density d of the Latin hypercube is set to 2000 and $\beta = [10, 10]^T$ is selected as discretization vector for the outputs. The probability density functions that are estimated for the different parameters by our method are depicted in figure 3, in comparison to the real density functions. As can be seen in the figure, the probability density functions of a_4, b_2, and b_3 are approximated very well. Also for a_1 we can see good agreement. The distributions peak at approximately the same point and the shape is roughly the same.

Although all parameters are identifiable, there are huge differences in the identifiability of the single parameters. The cumulative probability functions of a_4 and b_3 can be estimated very well with a sampling density d of only 250 (results not shown). For the approximation of b_2 and
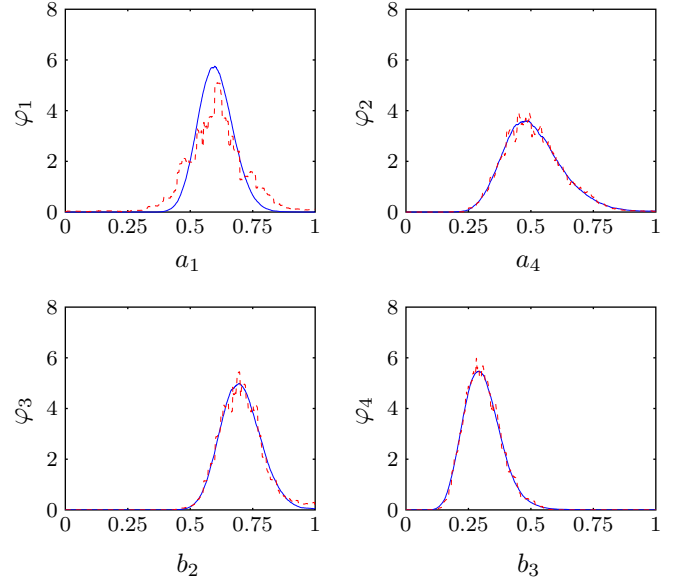

Fig. 3. Comparison of actual (dashed) and estimated (full)
parameter probability function.

especially a_1 more samples in the parameter space have to be taken. This can be related to the observation from the analysis done in Section 4.4, that a_4 and b_3 are of high relevance for the bimodal response of the cells, while the other two parameters do not have a high influence on this property.

For the considered example, the computation is quite efficient. Computation time on a standard desktop computer was on the order of a few minutes, and most of the time was spent computing trajectories for individual parameter values. In the proposed algorithm, this task can easily be parallelised for more complex models.

4.4 Analysis of the population model

In this section, we discuss the biological conclusions that can be drawn from a computational analysis of the population model (17) with parameter distributions as used in Section 4.3. For the considered system, it is of particular interest to distinguish between cells that undergo apoptosis and cells that stay alive. In apoptotic cells, the state variable x_2 (caspase 3 activity) tends to a positive value larger than a threshold θ. For non-apoptotic cells, x_2 returns to zero after a small transient rise.

In order to investigate the underlying differences that lead to such a differential behaviour, we consider a sample of parameter values $p^{(i)}$ taken from the distributions specified in Section 4.3, giving rise to trajectories $x^{(i)}(t)$ of (17). The parameter samples are clustered into the apoptotic set $\mathcal{A}$ and the non-apoptotic set $\mathcal{L}$ by the criterium

$$\mathcal{A} = \{p^{(i)} \mid x_2^{(i)}(T_{end}) \geq \theta\}, \quad \mathcal{L} = \{p^{(i)} \mid p^{(i)} \notin \mathcal{A}\}, \tag{18}$$

where $\theta = 0.3$ for this study.

First let us compare the sets $\mathcal{A}$ and $\mathcal{L}$ by directly examining the respective parameter values. As seen from Figure 4, the differences between the cells can mainly be explained from differences in the value of the parameter b_3, which is the threshold for NF-κB to inhibit caspase 3 activation,
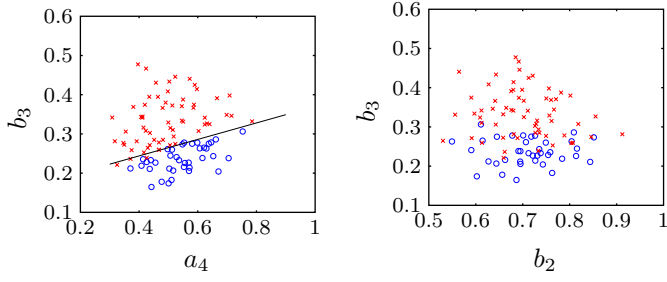
Fig. 4. Comparison of parameter values for apoptotic (x) and non-apoptotic (o) cells and approximate separation (19) in the a_4–b_3 plane.
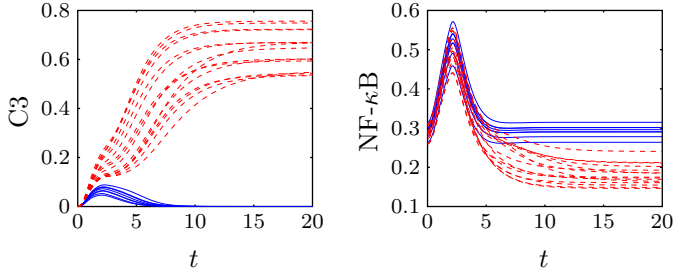

Fig. 5. Comparison of a few trajectories for apoptotic (dashed) and non-apoptotic (full lines) cells.

and the parameter a_4, which is the threshold for NF-κB to activate I-κB. In fact, an approximative separation criterium can be obtained directly from Figure 4 as

$$p \in \mathcal{A} \Leftrightarrow b_3 \gtrsim 0.16 + 0.21a_4. \qquad (19)$$

Apoptotic cells are thus characterised by high values for b_3 and low values for a_4, which relates well to biological intuition. The parameter b_2, which is the threshold for caspase 3 to inhibit NF-κB activation, seems to have little to no influence on the cell fate. These results indicate that the cell fate is determined by influences from the NF-κB pathway to the caspase cascade, and not vice versa.

Next, we try to find early-stage markers for the cell's fate. This is of interest because individual parameter values are not known when observing a single cell by e.g. live cell imaging, yet we may want to predict the fate of a specific individual cell. The collection of trajectories for the considered parameter sample is shown in Figure 5. Obviously, early-stage caspase activity is a good indicator for the later fate of the cell. However, it is quite interesting from the biological viewpoint that early-stage NF-κB activity seems not to be a good indicator. In fact, the NF-κB trajectories in Figure 5 separate only for $t > 10$, a time for which most apoptotic cells show already high caspase activity.

5. SUMMARY AND CONCLUSIONS

Heterogeneity in cell populations is an important aspect for research in systems biology. However, computational approaches to deal with heterogeneous populations are rare. A reasonable way to describe heterogeneity is to assume that parameter values are stochastically distributed within the population.

In the modelling process, it is then necessary to estimate the parameter distribution functions from suitable experimental data. For this paper, we assume that the output distribution in the cell population is measured at discrete sampling times. We present an optimisation-based approach to estimate parameter distributions from such measurements, which minimizes the prediction error based on a suitable sampling of the parameter space. With the suggested latin hypercube sampling, the approach scales well to systems with a high-dimensional parameter space.

We applied the suggested estimation method to artificial data for a model of TNF signal transduction. For the parameters where heterogeneity was assumed, our method gives good estimates of the parameter distribution function. The results thus indicate that those parameters are identifiable from the measurements used in this setup.

6. ACKNOWLEDGMENTS

REFERENCES

M. K. Al-Banna, A. W. Kelman, and B. Whiting. Experimental design and efficient parameter estimation in population pharmacokinetics. *J. Pharmacokin. Biopharm.*, 18:347–360, 1990.

S. V. Avery. Microbial cell individuality and the underlying sources of heterogeneity. *Nat. Rev. Microbiol.*, 4:577–587, 2006.

S. Boyd and L. Vandenberghe. *Convex optimization.* Cambridge University Press, Cambridge, UK, 2004.

M. Chaves, T. Eissing, and F. Allgöwer. Bistable biological systems: A characterization through local compact input-to-state stability. *IEEE Trans. Autom. Control*, 53:87–100, 2008.

A. Doucet, N. de Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice.* Springer New York, 2001.

T. S. Gardner, C. R. Cantor, and J. J. Collins. Construction of a genetic toggle switch in Escherichia coli. *Nature*, 403:339–342, 2000.

T. C. George, S. L. Fanning, P. Fitzgeral-Bocarsly, R. B. Medeiros, S. Highfill, Y. Shimizu, B. E. Hall, K. Frost, D. Basiji, W. E. Ortyn, P. J. Morrissey, and D. H. Lynch. Quantitative measurement of nuclear translocation events using similarity analysis of multispectral cellular images obtained in flow. *J. Immunol. Methods*, 311:117–129, 2006.

D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press, Cambridge, UK, 2003.

N. V. Mantzaris. 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, 2007.

J. T. Mettetal, D. Muzzey, J. M. Pedraza, E. M. Ozbudak, and A. van Oudenaarden. Predicting stochastic gene expression dynamics in single cells. *Proc. Natl. Acad. Sci.*, 103:7304–9, 2006.

W. E. Ortyn, D. J. Perry, V. Venkatachalam, L. Liang, B. E. Hall, K. Frost, and D. A. Basiji. Extended depth of field imaging for high speed cell analysis. *Cytometry A*, 71:215–231, 2007.

O. D. Perez and G. P. Nolan. Phospho-proteomic immune analysis by flow cytometry: from mechanism to translational medicine at the single-cell level. *Immunol. Rev.*, 210:208–228, 2006.

M. Stein. Large sample properties of simulations using latin hypercube sampling. *Technometrics*, 29:143 – 151, 1987.

H. Wajant, K. Pfizenmaier, and P. Scheurich. Tumor necrosis factor signaling. *Cell Death Differ.*, 10:45–65, 2003.